Philips Technical Review

HEATING BY HIGH-FREQUENCY FIELDS

I. INDUCTION HEATING

by E. C. WITSENBURG *).   621.364.156:621.785

The heating of a metal by means of an electric current induced inside it is by no means new; attempts to put this method into practical application date as far back as the beginning of this century. Shortly after 1920 induction heating with high-frequency current was introduced in the manufacture of radio valves. It was not until some years later (1926-1929) that further theoretical and practical research led to a better insight being obtained, upon which is based the rapid development of induction heating since 1940. Since the frequencies and the power required for this method of heating are the same as those used in radio transmitters, this development is due to the great progress made in radio engineering during the last 20 years. — Although the theory, which allows of the most favourable frequency being chosen and the apparatus suitably dimensioned, now no longer holds any secrets, the possibilities of application have not yet been fully investigated. The author mentions as an example a number of processes (surface hardening, annealing, soldering, melting) where induction heating has already opened up important perspectives for the metal-working industry.

A further article will deal with capacitive heating, which in a certain sense forms the complement of induction heating: whereas the latter is confined to conducting materials, capacitive heating, which is based upon dielectric losses, is applied to non-conducting or poorly conducting materials.

Heating is a treatment that has to be applied for innumerable processes in industry to give the product the properties required. In by far the majority of cases the heat needed is generated externally, either by burning or by some other chemical process, or by an electric current flowing through a separate resistance wire or forming an arc. The heat is carried to the point to be heated by conduction, convection and/or radiation.

Apart from this indirect heating, however, direct heating is also possible, the heat then being generated inside the material to be heated. If this material is electrically conductive (metal, graphite) the heat can then be generated by sending an electric current through it. Mostly, however, the shape of the object (or of the part of it to be heated)

does not lend itself to connection to a source of electric current. Some of the few examples of such are resistance-welding and the heating to which the filaments of incandescent lamps and radio valves are submitted during their processing. In many cases the only practical way to generate the heating current in the material is the method of induction, thus with the aid of an alternating magnetic field. Hence the name given to the process to be dealt with in this article: induction heating.

Heat can be generated not only in conducting materials but also in insulating media, thanks to the phenomenon of dielectric losses. These losses arise when the material is situated in an alternating electric field, thus for instance between the plates of a capacitor connected to an alternating voltage. Hence the name capacitive heating.

Thanks to these two methods of heating many

*) Philips Telecommunication Industry, formerly N.S.F., Hilversum (Netherlands).

cases of thermal treatment in industry have been made possible which by a more conventional method would be less efficient or even impossible. The improvement in quality and increased production reached in many cases with induction or capacitive heating have already led to extensive application of both these methods, whilst it is to be expected that this industrial field of application will be considerably expanded in the years to come. Experience will then show more clearly which thermal treatments lend themselves best to induction or to capacitive heating and which can best be carried out by the means hitherto more commonly applied, since, as with every process, induction and capacitive heating methods naturally have their limitations too.

Apart from the fact that in induction heating as well as in capacitive heating the heat is generated *inside* the material, these two methods have another point in common, in that both with capacitive heating and also in by far the most cases with induction heating satisfactory results are only possible by employing h i g h f r e q u e n c i e s, which can only be obtained in a suitable manner by the methods of radio engineering. Otherwise there is such a great difference between induction and capacitive heating — if only by reason of the nature of the object that can be treated (conductors as opposed to non-conductors) — that they are best dealt with in separate articles. Induction heating will be dealt with first.

Heating processes which have in many cases been greatly improved by the introduction of high-frequency inductive heating are: h a r d e n i n g, a n n e a l i n g, b r a z i n g and s o l d e r i n g, and m e l t i n g. Still further possibilities are being sought in all directions, as evidenced by the extensive literature being published nowadays on these subjects. The great interest being taken in this matter might give the erroneous impression that this is an entirely new technique. However, already shortly after 1920 the Philips factories at Eindhoven, among others, were applying induction heating on an industrial scale for the heating (degassing) of the electrodes of radio valves. Since 1935 the same works have been using an induction furnace [1] for melting magnet steel and other alloys. In both these cases the high-frequency current is generated by means of a valve oscillator.

For certain applications, however, lower frequencies suffice, in which case the power may alternatively be supplied by rotary generators.

Some brief theoretical considerations may serve to explain the characteristic results to be achieved with induction heating. Before proceeding to discuss some of its applications a very simple example will therefore be given to elucidate the most typical phenomena.

General comments on induction heating

Let us consider what happens when a wire is wound round a steel spindle of say 2 cm diameter (*fig. 1*) and the ends of the coil formed by these turns of wire (called the work coil) are connected
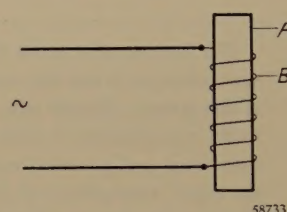
Fig. 1. A rod *A* is heated by the current induced in it by the alternating current in the work coil *B*.

to a generator supplying an alternating current with a frequency of say 500,000 c/s [2]). The current flowing through the work coil sets up a magnetic field which induces currents in the rod. Due to the I^2R losses these currents supply the heat required. In accordance with the known law of induction the currents flowing in the rod counteract the magnetic field from which they originated. From this it follows, as the first typical property of induction heating, that t h e h e a t g e n e r a t e d i s m a i n l y c o n f i n e d t o t h a t p a r t o f t h e w o r k p i e c e immediately opposite the work coil. Thus, given suitable dimensions and position of the coil, a workpiece can be p a r t i a l l y heated, and in many cases this will be found to be one of the main advantages of this method of heating.

There is also another special point about the induced current. The current density is greatest at the surface of the rod, rapidly diminishing inwards approximately according to the formula

$$J_x = J_0 e^{-\frac{x}{\delta}},$$

in which J_x represents the current density at a

---

[1]) For a description of this installation see Philips Techn. Rev. 1, 53-59, 1936.

[2]) A similar case, but not with the object of heating, was recently dealt with in this periodical: P. Zijlstra, An apparatus for detecting superficial cracks in wires, Philips Techn. Rev. 11, 12-15, 1949 (No. 1).

depth x, J_0 the density at the surface and δ the so-called penetration depth, i.e. the depth at which $J_x = J_0/e$. About 83% of the total heat is generated in the outermost layer of the rod of the thickness δ.

The depth of penetration depends upon the specific resistance ϱ and the relative permeability μ_r of the material of the rod and also upon the frequency f. It can be calculated with the aid of the formula

$$\delta = \frac{1}{2\pi} \sqrt{\frac{\varrho \cdot 10^7}{\mu_r f}} \approx 503 \sqrt{\frac{\varrho}{\mu_r f}} \text{ meters .} \quad (1)$$

(ϱ in $\Omega \cdot$m, f in c/s.)

As regards the relative permeability of ferromagnetic materials it is to be noted that on account of the very high field strength usual with inductive heating there is considerable saturation, so that the value of μ_r is always rather small (in the order of 100 or even 10). When the material is heated to above the Curie point (for iron 770 °C) the ferromagnetic properties disappear and the value of $\mu_r = 1$.

In fig. 2 the penetration depth has been plotted as a function of the frequency for several materials. From this it is to be seen that in the case of the rod in question (iron in the hot state) the heat is generated mainly in a skin 0.85 mm thick (at the chosen frequency of 500,000 c/s.)
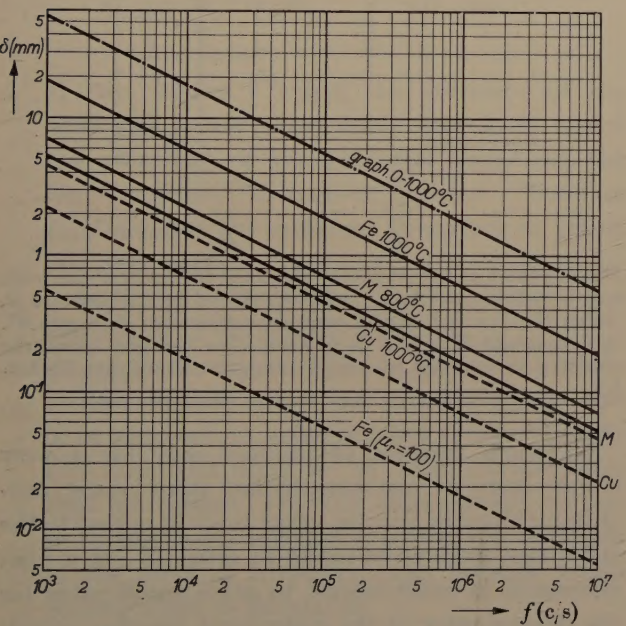

Fig. 2. The penetration depth δ as a function of the frequency f for several materials, according to formula (1). Graph = graphite, Fe = iron, Cu = copper, M = brass. The dotted lines apply for room temperature, the others for 800-1000 °C. The line for graphite is practically independent of the temperature. In the case of cold iron it has been taken that $\mu_r = 100$; for iron at 1000 °C $\mu_r = 1$.

This heat is used in the first place for heating the skin, but as soon as the temperature of the skin rises above that of the surroundings heat begins to flow away in two directions: through the rod inwards by conduction, and to a less degree outwards by radiation and convection. Mathematically it is difficult to determine exactly what happens, but from the heat conductivity and the coefficient of radiation of the material (in this case steel) an idea can be formed of magnitude of the two thermal flows. For instance, in iron a temperature difference of 100 °C per mm gives rise to a thermal current of about 7 W/mm². And at a temperature of 1000 °C at the surface the radiation amounts to 0.15 W/mm². We shall revert to the matter of heat dissipation when dealing with surface hardening.

Another point to be noted is that, if the shape of the workpiece makes it desirable, the work coil can also be placed inside the workpiece, an example of which will be given presently.

Efficiency of the work coil

Part of the power supplied to the work coil is lost owing to the resistance of the coil. Thus we can speak of the efficiency of the work coil as representing the ratio η of the power P_2 reaching the charge with respect to the power supplied $P_0 = P_1 + P_2$, with P_1 denoting the power lost in the work coil. Thus

$$\eta = \frac{P_2}{P_1 + P_2}. \quad \cdots \cdots \quad (2)$$

It is not difficult to understand that, within certain limits, the efficiency will be greater according as the specific resistance (ϱ_1) of the charge is higher with respect to that (ϱ_2) of the material of which the work coil is made, and also according as the charge is more closely enveloped by this coil. In addition to these factors there is another less obvious one bearing upon the efficiency of the work coil, namely the ratio of the diameter of the charge to the penetration depth δ_2 in the rod. A simplified theoretical calculation yields the following formula for the efficiency of the work coil:

$$\eta = \frac{1}{1 + \dfrac{D^2}{d^2}\left(1 + 6.25\dfrac{\delta_2^2}{d^2}\right)\sqrt{\dfrac{\varrho_1}{\mu_r \varrho_2}}}, \quad (3)$$

in which D = the diameter of the work coil and d = the diameter of the charge.

This formula can be derived in the following way [3]).

In the equivalent circuit (*fig. 3*) R_1 and L_1 represent the resistance and the self-inductance of the work coil, and R_2 and L_2 the resistance and inductance of the charge, the latter being coupled to the work coil via the mutual inductance M.
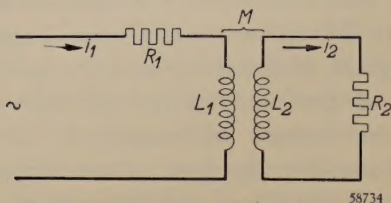
Fig. 3. Equivalent circuit for the situation illustrated in fig. 1. R_1, L_1 = resistance and inductance of the work coil, R_2, L_2 = resistance and inductance of the workpiece, M = mutual inductance.

If I_1 is the R.M.S. value of the primary current and I_2 that of the secondary current then the output is:

$$P_2 = I_2^2 R_2 = n^2 I_1^2 R_2,$$

where $n = I_2/I_1$ represents the current-transformation ratio.

The power P_1 lost in the work coil is

$$P_1 = I_1^2 R_1,$$

so that the efficiency of the work coil is:

$$\eta = \frac{P_2}{P_1 + P_2} = \frac{n^2 R_2}{R_1 + n^2 R_2} = \frac{1}{1 + \dfrac{R_1}{R_2} \cdot \dfrac{1}{n^2}} . \quad \cdots \quad (4)$$

We shall now consider separately the factors R_1/R_2 and $1/n^2$ occurring in the denominator of the last fraction, for a simple case taken very schematically.

Let us assume that the work coil consists of one turn of strip with such a width (H in *fig. 4*) that the field inside this coil can be regarded as being homogeneous. A rod-shaped workpiece (length $h < H$) is placed inside this homogeneous field.

The current in the work coil can be imagined as being concentrated in a thin layer (thickness δ_1) on the inside. The resistance of the work coil is therefore $R_1 = \varrho_1 . \pi D/H\delta_1$. Similarly, the resistance of the rod is $R_2 = \varrho_2 . \pi d/h\delta_2$.

For the sake of simplicity it is further assumed that the rod is s i m i l a r to the work coil, thus that $d/h = D/H$. In that case

$$\frac{R_1}{R_2} = \frac{\varrho_1}{\varrho_2} \cdot \frac{\delta_2}{\delta_1} .$$

Using formula (1) and taking $\mu_r = 1$ for the relative permeability of the work coil, consisting of non-ferromagnetic material, we arrive at

$$\frac{R_1}{R_2} = \sqrt{\frac{\varrho_1}{\varrho_2 \mu_r}} . \quad \cdots \cdots \cdots \quad (5)$$

There now remains the factor $1/n^2$. For the secondary circuit (fig. 3) we have:

$$i_2 (R_2 + j\omega L_2) + i_1 \cdot j\omega M = 0,$$

where i_1 and i_2 represent the complex values of the two

[3]) See also: E. C. Witsenburg, High-frequency inductive heating, (Dutch) T. Ned. Radiog. **12**, 201-211, 1947.

currents and $\omega = 2\pi f$. Thus

$$\frac{i_1}{i_2} = \frac{R_2 + j\omega L_2}{-j\omega M} .$$

$1/n^2$ is the square of the absolute value of this expression:

$$\frac{1}{n^2} = \frac{R_2^2 + \omega^2 L_2^2}{\omega^2 M^2} = \frac{\dfrac{R_2^2}{\omega^2 L_2^2} + 1}{k^2 \dfrac{L_1}{L_2}} .$$

in which $k^2 = M^2/L_1 L_2$, for which we can write approximately d^3/D^3.

Substituting:

$$R_2 = \varrho_2 \cdot \pi d/h\delta_2,$$
$$L_2 = \pi\mu F d^2/4h,$$
$$L_1 = \pi\mu F D^2/4H$$

(in which $\mu = \mu_0 \mu_r = 4\pi \cdot 10^{-7} \cdot \mu_r$ and F is a factor depending upon $D/H = d/h$) we find after some conversion:

$$\frac{1}{n^2} = \frac{D^2}{d^2} \left(1 + \frac{\varrho_2^2 \cdot 10^7}{\pi^2 F^2 f^2 \mu_r^2 d^2}\right) = \frac{D^2}{d^2} \left(1 + \frac{4}{F^2} \cdot \frac{\delta_2^2}{d^2}\right). \quad (6)$$

Substituting (5) and (6) in (4) and with $F = 0.8$ (corresponding to $D/H = 0.6$), we arrive at the equation (3) for the efficiency η.
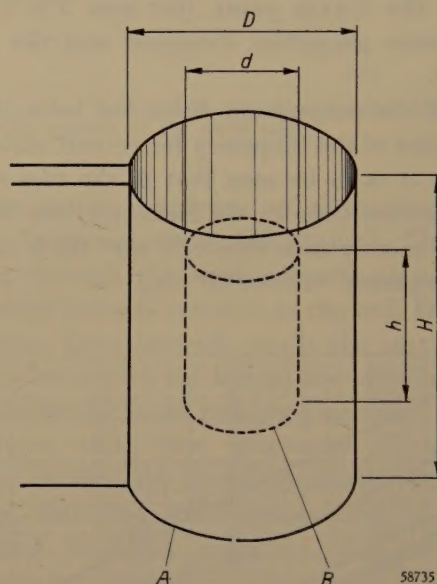
Fig. 4. Sketch of the set-up for which the efficiency of the work coil is calculated. A = work coil of one turn (diameter D, width H), B = rod-shaped workpiece (diameter d, length h). It is assumed that H/D is so great that the field inside the coil may be regarded as being homogeneous, and that $H/D = h/d$.

Although formula (3) applies only for a very simple case, we can nevertheless draw some conclusions from it also for other cases in regard to the steps that have to be taken in order to get a high efficiency, in other words, to get a small value for the denominator of the right-hand member of (3). The second term in this denominator consists of three factors, which will be dealt with separately.

1) D^2/d^2. For high efficiency the difference between D and d has to be chosen as small as possible; hence the work coil has to be made to fit as

closely as possible round the workpiece (or, inversely, the workpiece has to fit as closely as possible round the coil).

2) $(1 + 6.25\ \delta_2^2/d^2)$. Just as is the case with the first factor, this is also greater than 1. Allowing as the maximum value that can still usually be reached, say, 1.1, then we must have $d/\delta_2 \geq \sqrt{62.5} \approx 8$. In other, words, the frequency has to be chosen so high that the penetration depth is not greater than about one eighth of the diameter of the workpiece. Substituting for δ_2 the value according to (1), we get the following equation for the minimum frequency f_{min} required to reach a high efficiency:

$$f_{min} = 503^2 \cdot 62.5 \cdot \frac{\varrho_2}{\mu_r d^2} = 16 \cdot 10^6 \cdot \frac{\varrho_2}{\mu_r d^2}\ \text{c/s} \quad (7)$$

(ϱ_2 in $\Omega\cdot$m, d in m).

Taking as an example the case of iron heated to 1000 °C, with $\varrho_2 = 1.4 \times 10^{-6}\ \Omega\cdot$m and $\mu_r = 1$ (1000 °C is above the Curie point), then $f_{min} \approx 22.5/d^2$. From this we find the following values for various diameters of the rod:

d	f_{min}
500 mm ($\approx$ 20 '')	90 c/s
100 mm ($\approx$ 4 '')	2 250 c/s
10 mm ($\approx$ 0.4'')	225 000 c/s
5 mm ($\approx$ 0.2'')	900 000 c/s

If one chooses $f > f_{min}$ the efficiency of the work coil is not appreciably increased, whilst the efficiency of the generator supplying the high-frequency current is reduced. Therefore f_{min} is about the optimum frequency at which one can work.

In *fig. 5* the frequency f_{min} is plotted as a function of d for several materials.

3) $\sqrt{\varrho_1/\varrho_2\mu_r}$. As regards this factor, with a given charge we only know the value of ϱ_1. As was to be expected, in order to reach a high efficiency the work coil must be of high conductivity, thus of copper. For the same reason it has to be kept cold, and that is why it is often made in the form of a copper tube through which cooling water is circulated.

When the workpiece is likewise of copper and is still cold then the factor $\sqrt{\varrho_1/\varrho_2\mu_r} = 1$ and the efficiency remains less than 50%. In most cases, however, $\varrho_2\mu_r$ is much greater than ϱ_1, and higher values of η are therefore reached, even as much as 96%, as shown by *fig. 6*. In fig. 6 η has been plotted

as a function of D/d on the assumption that the frequency equals f_{min}. For non-ferromagnetic materials with a positive temperature coefficient of the resistance, such as copper, the efficiency
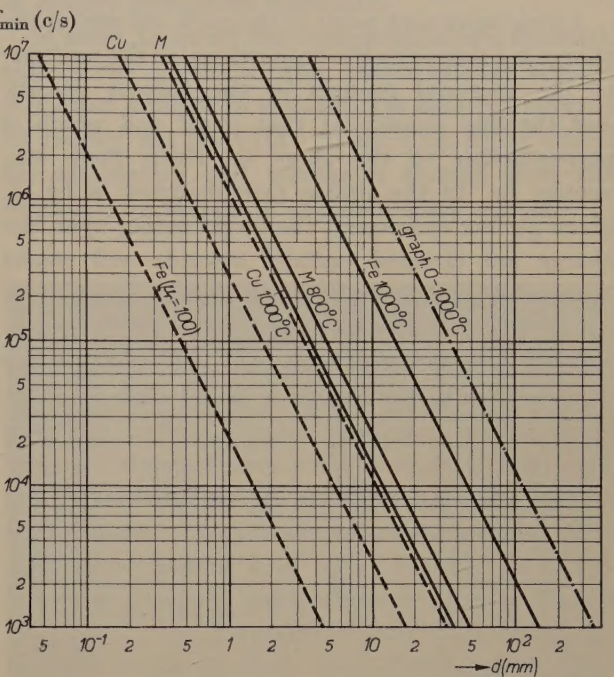

Fig. 5. The lowest frequency f_{min} at which the efficiency is not unnecessarily low, as a function of the diameter d of the workpiece, for graphite (*graph*), iron (*Fe*), copper (*Cu*) and brass (*M*), at room temperature (dotted lines) and at 800-1000 °C (fully drawn lines). For cold iron it has been taken that $\mu_r = 100$.

rises with the temperature. In the case of iron, on the other hand, the efficiency drops as soon as the temperature exceeds the Curie point, since μ_r then decreases more quickly than ϱ_2 increases.
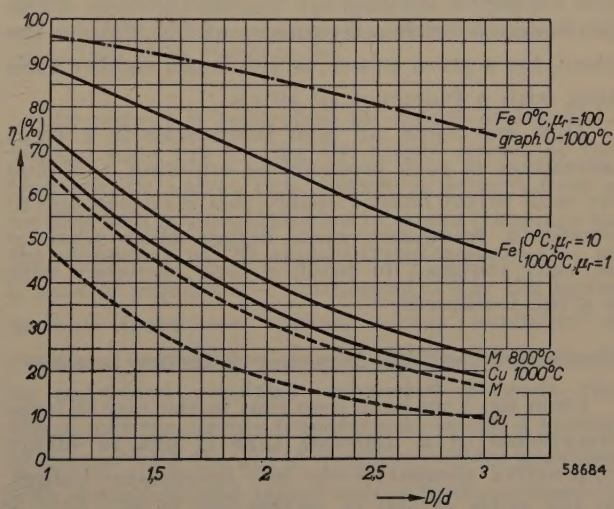

Fig. 6. Efficiency η of the work coil as a function of D/d (D = diameter of the work coil, d = diameter of the workpiece) for the same materials as in figs 2 and 5, at 0 °C (dotted lines) and at 800-1000 °C (fully drawn lines). For cold iron a curve has been plotted for $\mu_r = 100$ and another for $\mu_r = 10$. The frequency is $= f_{min}$.

Electrodynamic forces acting upon the charge

We shall conclude our general note with a consideration of the electrodynamic forces exercised upon the charge by the magnetic field.

Acting upon each point of the work coil and of the charge is a force which is proportional to the modulus of the vectorial product of the current density and the flux density at that point and which is perpendicular to the plane in which these two vectors lie. In the simple case sketched in fig. 1, given perfect symmetry, the forces acting upon the charge are directed radially inwards and cancel each other, so that the workpiece as a whole is not subject to any force. Where there are deviations from symmetry, however, such as occur, for instance, while the charge is being placed in the work coil, a (repelling) force is indeed exercised upon the workpiece and under certain circumstances this may prove to be troublesome. From what follows, where for the sake of simplicity the charge is assumed to be in a homogeneous field, it will be seen that this force becomes smaller as the frequency chosen is higher.

We have already seen that the depth of penetration of the current in the charge is proportional to $1/\sqrt{f}$ (see (1)), so that the resistance R_2 of the charge is proportional to $\sqrt{f}$. For a given power $I_2{}^2R_2$ in the charge $I_2{}^2$ must therefore be proportional to $1/\sqrt{f}$. The electrodynamic force acting upon the charge is proportional to the product of I_2 and the flux density B, so that if B is proportional to I_2 the force is proportional to $I_2{}^2$, thus proportional to $1/\sqrt{f}$; with a given power the force is inversely proportional to the square root of the frequency. Thus, for instance, with a frequency of 500,000 c/s the force, for a given power, is 100 times smaller than that with a frequency of 50 c/s.

This low force at high frequencies is a great advantage in such processes as soldering, to which we shall revert later. Possibly this may lead one to choose a frequency even higher than what is actually required to reach a high efficiency with a given workpiece.

Generators for high-frequency induction heating

From fig. 5 we have seen that in the case of large workpieces of a material that is not so highly conductive comparatively low frequencies suffice, in the order of 500 to 10,000 c/s. In the case of objects no more than a few millimetres in size, however, frequencies are required of the order of 1,000,000 c/s. Now fig. 2 shows that at the low frequencies just mentioned the penetration depth is comparatively great. Therefore, if for a certain heat treatment, particularly of iron, a small penetration depth is desired (e.g. 1 mm or less), as is often the case, then one should use a frequency in the order of 500,000 c/s.

The low frequencies can be generated by rotary converters of the type formerly used for wireless telegraphy on very long wavelengths. We shall not discuss these machines here, but a few words will be devoted to the valve generators with which, as is known, much higher frequencies can be generated.

The most important elements of such a valve generator are a triode and an oscillatory circuit. The circuit is caused to oscillate by means of feedback. The work coil is taken up in the self-inductance branch of the circuit, either direct or via a transformer. The latter is often necessary to bring about the optimum matching, i.e. to transform the very low resistance of the workpiece to a value at which the transmitting valve is correctly loaded.

At the resonant frequency the resistance r in series with one of the branches of an oscillating circuit (*fig. 7a*) is equivalent to a resistance R_p of the value Q^2r parallel to the circuit ($Q =$ quality factor). If, with a suitable value of Q (about 30), R_p has to be made of magnitude required for the transmitting valve (e.g. 1000 Ω) then r must be of
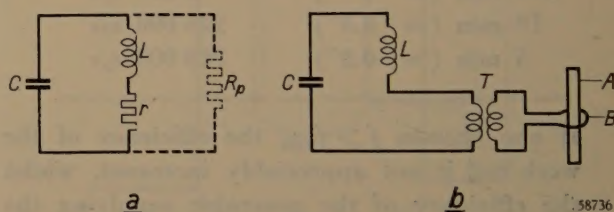

Fig. 7. a) At the resonant frequency of an oscillatory circuit L-C the resistance r in one of the branches is equivalent to a parallel resistance $R_p = Q^2r$ ($Q =$ quality factor).
b) The resistance of the workpiece A is often so low that an intermediate transformer T is required between the work coil B and the oscillating circuit L-C.

the order of 1 Ω. Now as a rule the resistance R_2 of the workpiece is much less than this. In the simple case considered in the foregoing $R_2 = \varrho_2 \cdot \pi d/h\delta_2$. Where $f = f_{\min}$, $d \approx 8\delta_2$ and thus $R_2 \approx 8\pi\varrho_2/h$. With, for instance, $h = 5$ cm we find for iron at a temperature of 1000 °C: $R_2 \approx 0.7$ mΩ. To transform this value to $r = 1\Omega$ a work coil with $\sqrt{r/R_2} \approx 38$ turns would be required, and usually this is not practicable. The desired transformation is then brought about with the aid of an intermediate transformer (T in fig. 7b), which then allows of a work coil being used consisting of only one turn. This results in the heating being highly concentrated and for that reason the transformer T is called a **concentrator**.

Since the resistance of most materials changes appreciably with the temperature, in order to maintain the optimum conditions repeated match-

ings would in fact have to be made in the course of the heating process. For melting processes, lasting say 10 minutes, this can be done by connecting capacitors in parallel to the work coil
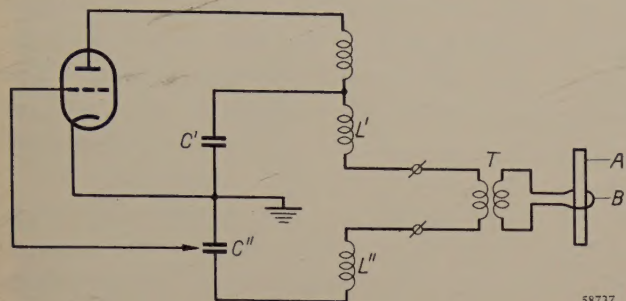

Fig. 8. Basic circuit diagram of a generator for high-frequency induction heating (Colpitts circuit). C'-C''-L'-L'' = symmetrically designed oscillatory circuit, T = intermediate transformer, B = work coil, A = charge.

step by step [4]). In many other cases, however, the heating process takes no more than a few

seconds, so that manual control is impracticable. Although automatic control would be possible, a fixed adjustment averaging approximately the optimum adjustments for the cold and the hot state is usually considered sufficient.

In order to avoid the complication of a feedback coil in the generator (for constructional reasons it is often difficult to use one) it is usually preferred to employ the Colpitts circuit (*fig. 8*), in which the alternating grid voltage is derived from the capacitive branch of the oscillatory circuit.

Fig. 10. Generator for high-frequency induction heating, type SFG 134/00. On the right panel the connections for the work coil, with water cooling. Maximum power in the workpiece 20 kW, frequency 450 kc/s. The apparatus is equipped with two valves TA 12/20 connected in parallel and three rectifying valves DCG 12/30. Height in mm.

The output is controlled by means of the D.C. voltage used for feeding the anode circuit of the transmitting valve. This voltage is derived from a rectifier and is variable either in stages (tappings on the secondary of the supply transformer) or — as is mostly required — continuously. For the latter method there are various means available: a variable transformer (with sliding contact) or an induction regulator connected between the mains and the supply transformer, or else a system of control with the aid of rectifying valves provided with a control grid (relay valves or thyratrons) [5].

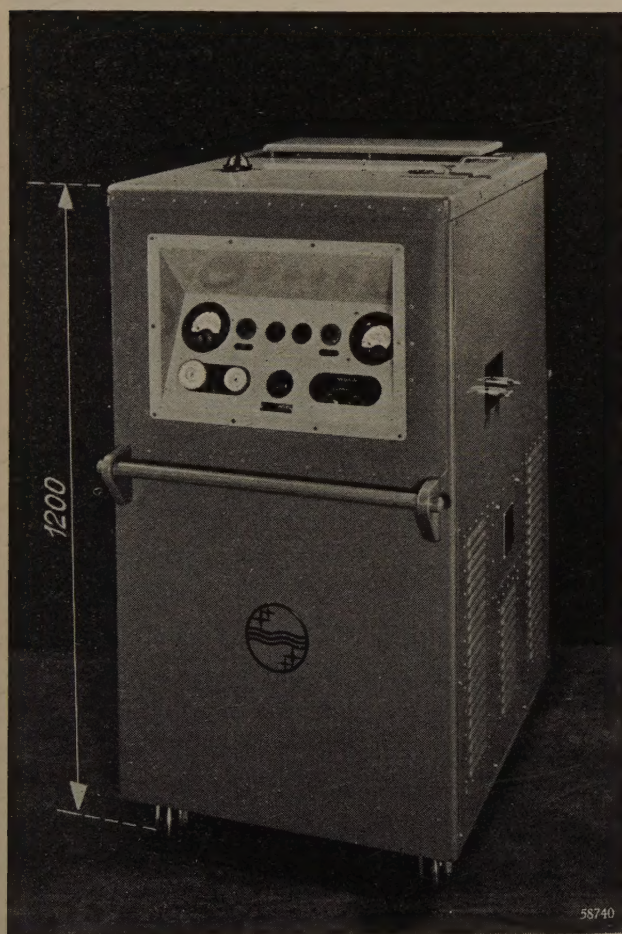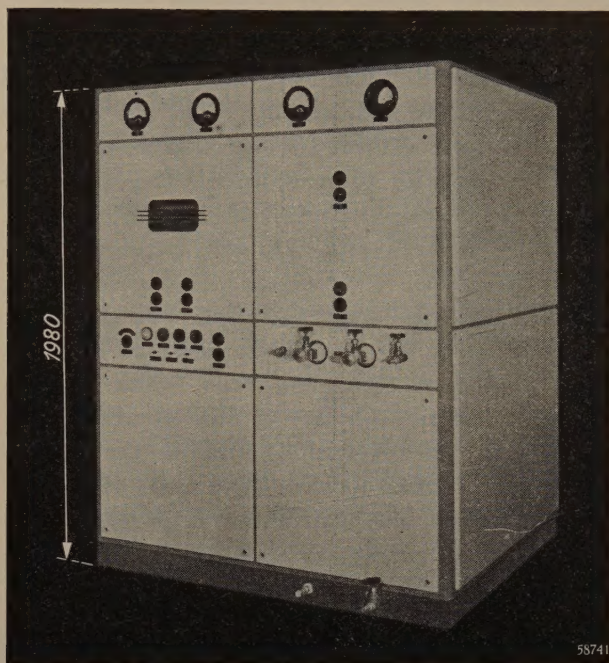*Figs.* 9 and 10 give photographs of generators for

Fig. 9. Generator for high-frequency induction heating, type SFG 136/00. On the right the connections for the work coil. Maximum power in the workpiece 2 kW, frequency 1.2 Mc/s. The apparatus is equipped with one transmitting valve TB 3/2000 and two rectifying valves DCG 5/5000. Height indicated in mm.

[4]) See page 55 of the article quoted in footnote [1]).

[5]) See, e.g.: J. G. W. Mulder and H. L. van der Horst, A controllable rectifier unit for 20,000 V and 18 A, Philips Techn. Rev. 1, 161-165, 1936.

induction heating, for powers of 2 kW and 20 kW. These generators have a fixed frequency of 1.2 and 0.45 Mc/s respectively, which according to fig. 5 are high enough for heating objects of a diameter down to a few millimetres with a reasonable efficiency and a small depth of penetration.

Applications of induction haeting

We shall now describe some applications of induction heating where this method yields exceptionally favourable results.

Surface hardening

Many parts of machinery, such as spindles and gear-wheels, and also cutting tools, for instance screw taps, are subject to wear and at the same time are subjected to great forces. The steel used for making these parts and tools can be hardened to give it better wearing properties, but this also makes it brittle. To withstand any great force, however, the steel must be tough, but then it is also soft. It is not possible to combine hardness with toughness throughout the whole of the material, but in fact this is not essential. It is sufficient if the material is hard on the surface where it is subject to wear, and everywhere else tough. To reach this distribution of properties the case-hardening process is traditionally employed: the workpiece is made of a non-hardening steel and the surface layer is subjected to the action of carbon, thus forming in that layer a steel that can be hardened. For this case-hardening a lengthy process of annealing is essential, and in that process undesired alterations may take place in the structure of the material and there are apt to be deviations from the specified dimensions.

A much better solution of the problem of surface hardening is offered by the process of high-frequency induction heating. With this method heat is generated in a layer on the surface, just where it is needed. It is now only necessary for the outermost layer of the non-hardened, tough carbon steel (hardness about 15 Rockwell) of the work-piece to be heated to the hardening temperature so quickly that the deeper layers, receiving heat only through conduction, are kept at a much lower temperature, after which the outermost layer is rapidly cooled. This layer then has a hardness of about 60 Rockwell. In the process of heating the power supplied has to be raised high enough to exceed amply the dissipation by conduction. This means that the surface density of the power supplied must be of the order of

1 kW/cm². Such powers and even higher are actually employed in induction heating (the highest values correspond to the thinnest hardened layers).

No other means are known for reaching such a high concentration of heat, excepting the electric arc as used for welding, but this cannot be sufficiently controlled for the hardening of metals. The transmission of heat by radiation or direct contact with, for instance, an acetylene flame is absolutely inadequate.

The concentration that can be reached in practice with induction heating is limited by the losses in the work coil. Therefore in order to get very high concentrations one must take into account the already discussed factors governing the efficiency
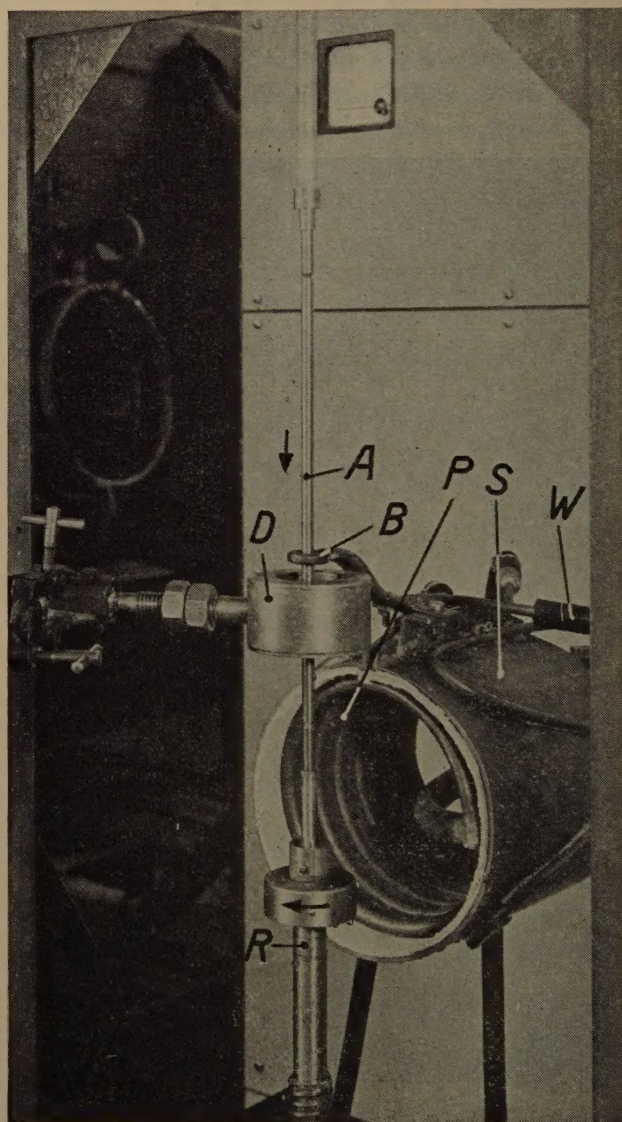
58680

Fig. 11. Hardening of a carbon-steel spindle (A). B = work coil of one turn, P = primary, S = secondary coil of the intermediate transformer (T in fig. 8), D = water spray for quenching the spindle, moving downwards at a rate of 2 m/min. and at the same time rotated by the shaft R. W = cooling-water supply for the work coil.
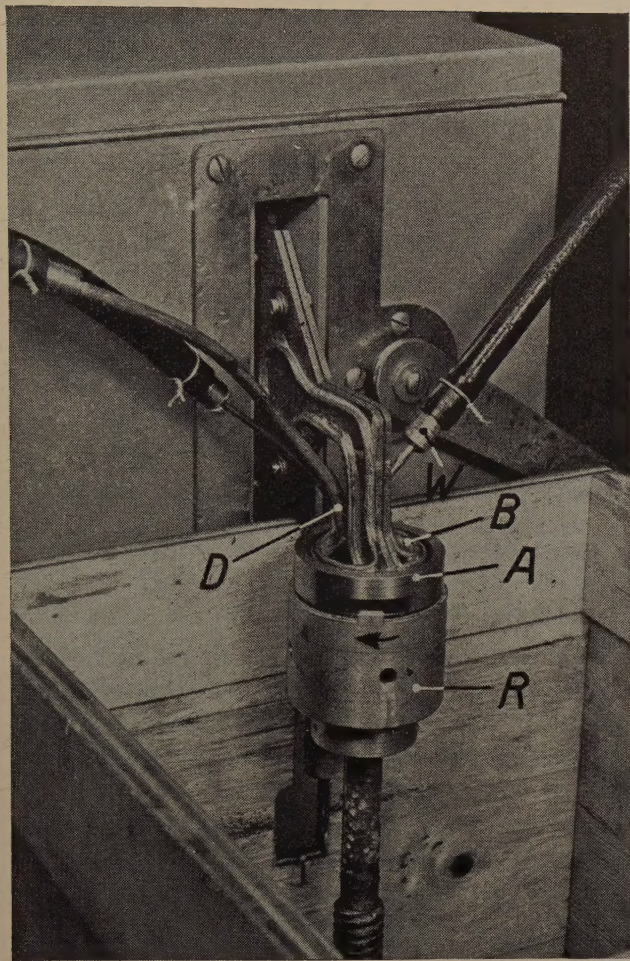
58681

Fig. 12. Hardening the inside of a carbon-steel ring (A). The letters B, D, R and W have the same meaning as in fig. 11.

of the work coil and arrange for this coil to be well cooled with water (so as to carry off the loss heat and thus keep the resistance of the coil low).

The process of surface hardening will now be illustrated with three examples. *Fig. 11* shows a set-up for hardening small spindles (6 mm in diameter) made of carbon steel. These are passed through a work coil at an axial speed of 2 m/min. Immediately underneath this coil is a water-spray for quenching. During this process the spindle is rotated in order to neutralize the effect of small asymmetries.

After the treatment the spindles have an outer layer 0.5 mm thick with a hardness of 64 Rockwell, whilst the hardness of the core remains unchanged. Thus the core retains its toughness, and it is due to this that the spindle shows very little warp (owing to inhomogeneity of the base material): it is at most 0.07 mm over a length of 300 mm.

The set-up of *fig. 12* is used for hardening a ring of carbon steel on the inside. The work coil is lowered into the ring from above. The ring lies on a rotating bush, again to neutralize asymmetry. The treatment is carried out with a 20 kW plant and takes no more than a few seconds. The result is a layer of 0.8 mm with a hardness of 60 Rockwell, whilst at 2 mm depth the hardness is 15 Rockwell.

The experimental set-up of *fig. 13* serves for the hardening of endless band-saws, which are passed under the work coil at a rate of 2 m/min. This
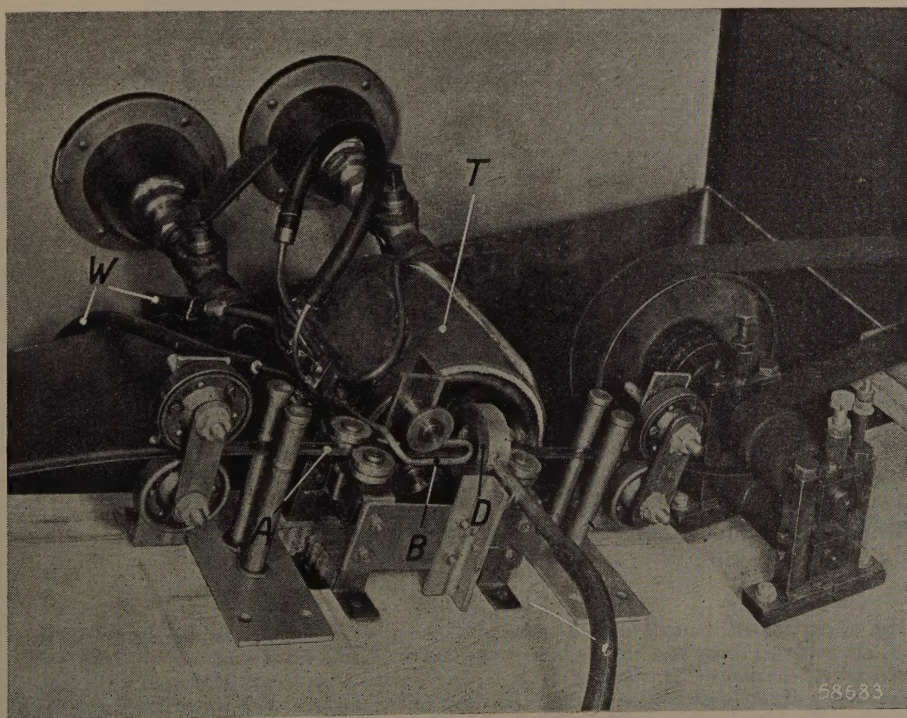
58683

Fig. 13. Hardening a band-saw passed under the coil B at a rate of 2 m/min. T = concentrator; meaning of the other letters as in fig. 11.

coil is so shaped that both the toothed and the plain sides of the band are heated, to avoid warping.
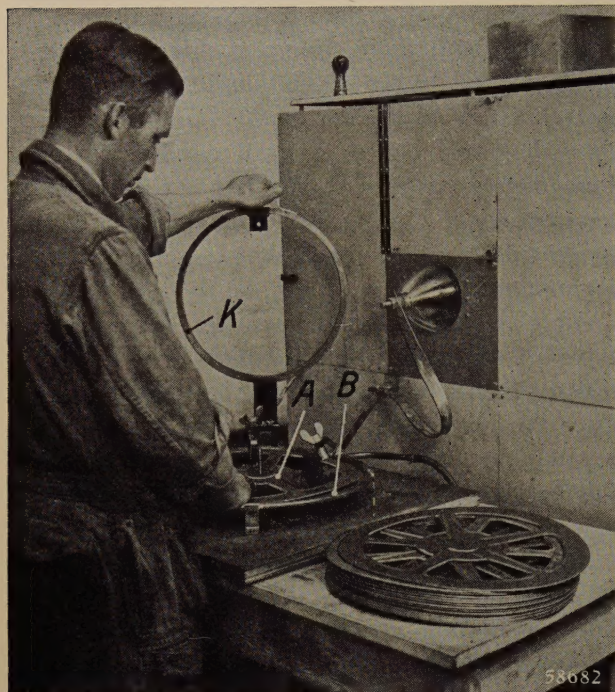

Fig. 14. Removing stresses from the rim of punched discs (flanges of film spools) by annealing. A = disc about to be treated. B = work coil. K = hinged ring for holding the disc.

Along the edge a hardness of 64 Rockwell is reached, at 1.5 mm away from the edge 60 Rockwell, whilst in the middle the hardness of 15 Rockwell remains unchanged.

Annealing

In the case of processes like forcing and deep-drawing, which mostly take place in several steps, often the workpiece becomes too hard and has to be annealed. Particularly when this has to be done locally, the method of induction heating has its advantages. As an example may be mentioned the beading of the rim of reflectors. *Fig. 14* gives another example, where the rim of punched discs (flanges of film spools) is inductively heated for 7 seconds to neutralize the stresses due to the punching.

Brazing and soldering

Induction heating can be carried out so quickly that hardly any oxidation of the workpiece arises. This makes this method highly suitable both for brazing and for soldering, for which purposes it is in fact already being applied on a large scale. Brazed joints requiring several minutes' heating with a flame can now, thanks to the method of induction heating, be made in a few seconds.
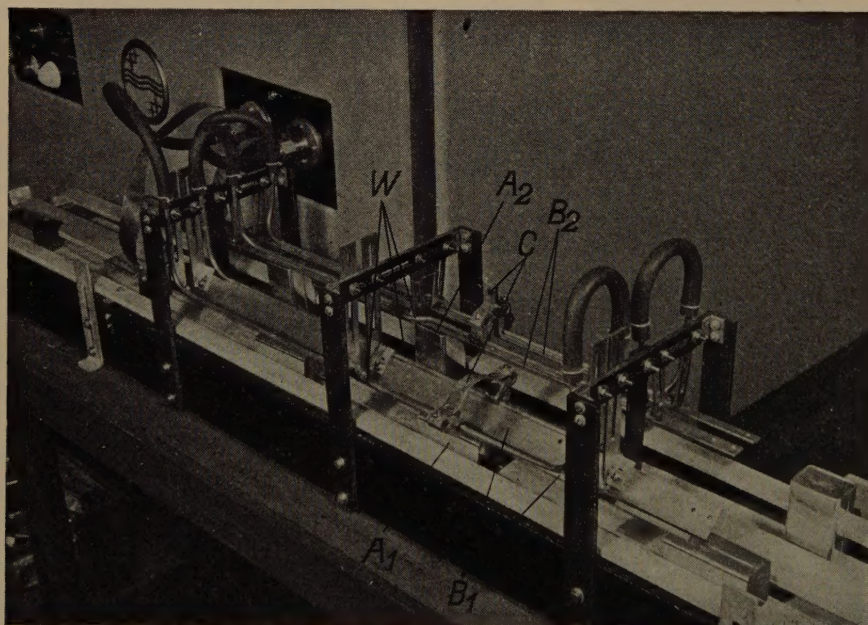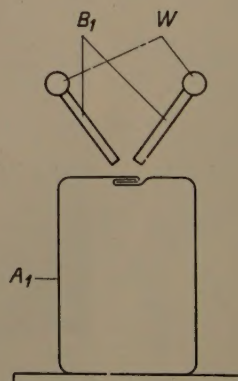
Fig. 15. *a*) Soldering small tins fed underneath the work coil by a moving belt. The tins A_1 are soldered along the longitudinal seam and at A_2 the lid is soldered on. The work coil for A_1 consists of the slanting strips B_1 (see the cross section in sketch *b*), that for A_2 consisting of the horizontal strips B_2. The work coils are cooled with water flowing through a tube (W). The slanting position of B_1 gives a better distribution of current (concentrated at the narrow sides of the strips facing the workpiece). The length of the work coils can be adjusted with movable clamps C.

Another point of importance in a process like brazing or soldering is the already mentioned fact that the forces acting upon the workpiece are small, provided a sufficiently high frequency is used. It is then a very simple matter to place the workpiece inside the coil; often it can be done by hand, or the workpieces can be fed into the heating apparatus by a moving belt without — although they may be light — being displaced by the electrodynamic forces.

Fig. 15 shows a set-up for the soldering of small tins. For particulars see the explanation given in the caption.

Melting

Owing to the short melting times that can be reached with induction heating (e.g. 10 to 15 minutes) this method is excellently suited for the melting of alloys where an exact composition is of importance. With such a short melting time the melt has little opportunity to be affected by the atmosphere and any volatile components have little chance of escaping.

In the article referred to in footnote [1]) a description is given of the plant used at Eindhoven for melting magnet steel and other alloys. The capacity of this plant is 200 kg steel per hour. The trans-

mitting valve (type TA 20/250) has an output of 250 kW and works at a frequency of 5000 to 10,000 c/s. Similar installations have meanwhile been taken into use elsewhere.

These examples should suffice for the present, but, as already stated, the possibilities of induction heating are by no means exhausted and new applications are being sought in all directions. It can therefore be taken as a certainty that more and more use will be made of this new technique in industry.

———

Summary. A feature which induction heating has in common with capacitive heating is the fact that the heat is generated inside the object, contrary to the case with most of the other methods of heating. A formula is derived, for a simple case, for determining the efficiency of the work coil through which the current passes that induces the heating current in the workpieces. From this formula it appears, inter alia, that in order to work with a good efficiency the frequency has to be chosen sufficiently high to ensure that the penetration depth in the workpiece is not greater than about one eighth of the diameter of the workpiece. This small depth of penetration confines the generation of heat to a thin layer on the surface, a fact which is turned to good account in many applications. It is also shown that with a given power the electrodynamic forces acting upon the workpiece are smaller according as the frequency chosen is higher. After a brief discussion of valve generators for induction heating, some applications are dealt with: surface hardening, annealing, brazing and soldering, and melting.

LOT INSPECTION BY SAMPLING

by H. C. HAMAKER.

If for some reason the reader should be interested in the number of printer's errors in this journal he would not be likely to go punctiliously through all the pages, but he would take a sample. Sampling procedures are in fact quite common; everyone uses them, often without much thought. Where, however, mass-production processes have to be checked by means of samples, large sums of money may be involved in carrying out the sampling procedure and in utilizing its results. Hence it is of value to obtain an insight into the various factors playing a part in the organization and operation of such lot-by-lot inspection tests.

In the mass-producing factory we are often faced with the problem of testing an inspection lot [1] of components or of finished articles in order to decide whether they come up to specification and thus may serve their intended purpose. In a great many such cases it is not economically justifiable to inspect every item in the lot, so that one has to resort to the taking of a sample. The question then arises how this sample should be taken, how large its size should be, and what requirements it should satisfy, in order that it can be decided with a reasonable degree of certainy whether the inspection lot as a whole is "good" or "bad". Some fundamental aspects of this problem will be discussed below.

We shall confine our considerations to inspection lots consisting of a number of concrete units, as for example an inspection lot of screws, resistors or incandescent lamps, which are inspected according to "attributes"; that is each item is classified as either "good" or "bad", and it is accordingly "accepted" or "rejected". An item upon inspection found to be defective is called a "reject". The percentage of defective items contained in an inspection lot is called the "percentage defective".

Each inspection lot contains a certain unknown percentage defective, and it is desired to provide from a sample some useful information concerning this percentage. The problem is how this can best be done.

Fundamentals; the operating characteristic

To cast our arguments in a concrete form let us consider a practical example. We shall suppose that we have received from an outside supplier a lot of, say, 100,000 small rivets. A sound riveted joint requires that a rivet fits properly into its hole, and consequently some trouble will be experienced when the rivet is burred. Let us assume that if the lot contains not more than 5% rivets with burs this is not serious, but that a higher percentage causes an undesirable delay in the riveting work. It is therefore specified that to be accepted a lot may not contain more than 5% of burred rivets. Whether this condition is satisfied or not has clearly to be decided on the basis of a sample, because inspecting each rivet separately would require an unwarranted amount of labour.

The simplest procedure is that of "single sampling": we take out of the lot a sample of n_0 rivets which we inspect to see how many of them are burred. The number n_0 is called the "sample size".

If reliable conclusions are to be drawn from the sample concerning the lot as a whole, we have first of all to see to it that the sample is taken with care. It is, for instance, conceivable that the rivets in the lot have come from different machines producing different percentages of burs. If the lot has been delivered in one sack it may well be that the percentage of burrs near the bottom is greater than in the top layers, so that a sample taken for the sake of convenience from the top alone is not representative of the lot as a whole. It is therefore of essential importance that the sample should be drawn from different levels. The ideal procedure would be to extract one rivet at a time, to shake the sack thoroughly before extracting the next one, and so on until the sample is complete. Mostly, however, this is not practicable and we shall have to approximate our ideal by taking, say, five or ten sub-samples from different spots and combine these to make up our final sample for inspection. In what follows it will always tacitly be assumed that such precautions have been taken.

---

[1] The term "inspection lot" or briefly "lot" will be used throughout this article for a collection of items accepted or rejected as a whole on the basis of a sampling plan.

Let us now suppose that in a random sample of 200 rivets 6 have been found to have burrs, that is 3%. In judging the lot on the basis of this datum two lines of argument are open.

Since the sample contains 3% of rejects the lot is obviously not free of defectives. It is also evident that the consignment is not likely to contain more than, say, 20% of defectives, for it would then be highly improbable that our sample contained only 3% of rejects. Thus we are clearly able to form some sort of estimate of the percentage defective in the lot. This percentage will lie somewhere around 3%, but as only a comparatively small sample has been inspected it may not be concluded that this percentage is exactly 3%. Within what limits, then, is it reasonable to suppose the percentage defective to lie and what confidence should we place on such a judgment?

Though mathematical statistics furnish a clear answer to this question, we shall not consider this aspect of our problem in detail, because it does not directly bear on the problems encountered in the factory. There it is desired to lay down clear rules such as the following: take a sample of size n_0 accept the inspection lot when it contains no more than c_0 defectives and reject the lot when this number is exceeded. The figure c_0 is commonly called the "acceptance number". Now for a proper understanding of the meaning and consequences of such a rule it is appropriate to look at the problem from a somewhat different angle than has been done above.

To this end we no longer ask what conclusions can be drawn from the sample regarding the unknown percentage defective in the inspection lot, but instead we enquire into the results that would be reached when an inspection lot with a prescribed percentage defective were repeatedly subjected to a certain "sampling procedure" (for example to a single sampling inspection specified by a sample size n_0 and an acceptance number c_0).

This can easily be settled by experiment. We take, for instance, some thousands of small pieces of cardboard, identical in shape, a small percentage of which has been marked as "defective" and then thoroughly mixed with the rest. From this artificially constructed inspection lot with known percentage defective we then proceed to draw a series of random samples in each of which the number of rejects is determined. The outcome of such an experiment has been recorded in *table I*, where the various columns refer to inspection lots containing different percentages of defectives.

Table I. a) The numbers of rejects observed in 10 successive samples taken from the same inspection lot. The figures in the various colums give the results for different percentages defective in this inspection lot. b) The result of judging by the sample when 6 is the maximum number of permissible rejects. c) The probability of acceptance with the sampling plan considered ($n_0 = 200$, $c_0 = 6$).

Percentage defective	1%	2%	3%	4%	5%	6%
a) Number of rejects in each sample	1	4	6	5	14	8
	1	9	5	6	14	11
	1	2	6	13	8	16
	1	3	7	6	9	6
	0	5	10	6	10	11
	3	8	9	8	9	9
	0	4	7	9	14	14
	2	6	15	2	11	8
	1	4	9	9	8	19
	2	6	4	7	5	14
b) Accepted	10×	8×	4×	5×	1×	1×
Rejected	0×	2×	6×	5×	9×	9×
c) Probability of acceptance	0.995	0.889	0.606	0.313	0.130	0.046

Tables I assumes, it is true, that ten samples were drawn from one and the same inspection lot, which is never done in practice. The same results would, however, have been obtained if the ten samples were drawn from ten different inspection lots each of them containing the same percentage defective, and, interpreted in this way, table I corresponds more closely to the conditions prevailing in the factory. In the course of time a large number of inspection lots will be submitted for inspection differing one from the other in the percentage of defectives they contain. Amongst these now and again inspection lots will occur with a percentage defective of 5% and the data recorded in the fifth column of table I may then be considered as the numbers of rejects observed in the successive samples drawn from these specific lots.

If, as we assumed above, inspection lots containing more than 5% of burred rivets are unacceptable, table I shows at once that it would be wrong to prescribe that there should be not more than 5% of rejects in the sample, or in other words that there should be no more than 10 rejects in a sample of 200. From the last column we see that, under such a specification, out of ten lots with 6% defectives four would still have been accepted, the number of rejects being less than 10. Consequently to satisfy our demands the requirements laid down for the sample should be more severe.

Let us therefore see what happens if only 3% of defectives are permitted in the sample, so that among 200 items inspected there should be no

more than 6 rejects. The results then obtained have been recorded at the bottom of table I. The ten inspection lots with 1% defectives are all accepted, two of those containing 2% defectives are rejected, and so on; with one exception all the lots with 5 or 6% are rejected, which answers our requirements much better than the previous case.

The vagaries of chance are also manifest: of the inspection lots with 3% defective six are rejected, whereas of those with 4% defective only five are rejected.

The whole procedure is very much like a lottery. For each inspection lot a draw is taken which leads to acceptance if it yields a prize, but leads to rejection if it is a blank. Evidently the chance of drawing a prize diminishes as the percentage defective in the inspection lot increases.

Theoretically these principles are expressed by assigning to each inspection lot a certain "probability of acceptance", P. In accordance with the theory of probability P is so defined that it is a number between 0 and 1. A probability of acceptance of 0.30 signifies that if the corresponding lot were repeatedly subjected to the sampling procedure it would be accepted on the average in 30% of the cases and would be rejected in 70%. Once the sampling plan — i.e. the detailed specification of the sampling inspection procedure — has been fixed the probabilities of acceptance can be computed by application of the theory of probability.
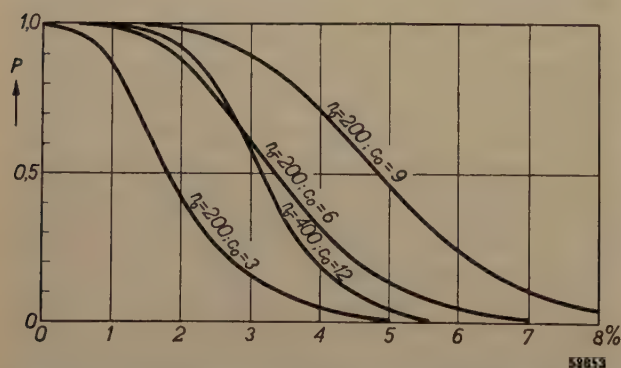

Fig. 1. Operating characteristic of some single sampling plans, i.e. the probability of acceptance plotted as a function of the percentage defective in the inspection lot; n_0 is the sample size, c_0 the acceptance number, or the number of rejects allowed.

For a single sampling plan with a sample size $n_0 = 200$ and an acceptance number $c_0 = 6$ the probabilities of acceptance thus obtained have been entered in the bottom row of table I. By plotting the probability of acceptance as a function of the percentage defective we obtain curves which are called the "operating characteristics"; some examples are represented in *fig. 1*. Each sampling inspection plan possesses its operating characteristics which are invariably curves of the same shape, though they may differ in location and slope.

If, keeping the sample size constant, we increase the acceptance number the operating characteristic shifts towards higher percentages, as would indeed be expected. For a sample size $n_0 = 200$ and an acceptance number $c_0 = 6$, 3% of rejects are permitted in the sample and the operating characteristic is seen to be steepest in the neighbourhood of this value. If we increase the size of the sample, while leaving the percentage of permissible rejects unaltered, the operating characteristic remains at about the same place but assumes a steeper slope, as is evidenced by the curve for $n_0 = 400$, $c_0 = 12$. In the extreme case, if the entire lot were subjected to inspection, we know the percentage defective with precision and we can decide with certainty whether it is higher or lower than the maximum value permitted. The operating characteristic then consists of a horizontal part at $P = 1$ for percentages defective less than the permitted maximum, and a horizontal line at $P = 0$, for higher percentages, these two horizontal segments being interconnected by a perpendicular located at the permitted maximum.

This is of course a purely theoretical case, because inspection of the entire lot would enable us to remove all the rejects it contains, so that we should be left with a lot without defectives. In this connection it should also be mentioned that a similar simplification has been introduced in plotting the curves in fig. 1, it having been assumed that the sample comprises only a small fraction of the inspection lot, so that the number of rejects detected and consequently removed from the sample can be ignored. These, however, are details which do not affect the general trend of our argument.

The purpose of these arguments was merely to show that the operating characteristic usefully portrays the practical performances of a sampling inspection plan. Hence these characteristics will play an important part in our further considerations.

Different sampling plans

So far we have spoken of sampling plans as the concrete specifications for taking a sample and acting upon the result of its inspection; we have only considered single sampling plans characterised by a sample size n_0 and an acceptance number c_0. All simple sampling plans will henceforth collectively be designated as the "single sampling system".

Apart from the single sampling system other

systems have been developed and successfully applied, the most important being the "double" and the "sequential" sampling systems. These possess the advantage of requiring on the average fewer observations than the single system, which is, however, partly offset by their being somewhat more complicated. Later on we shall return to the problem of choosing between these different methods, confining ourselves for the moment to explaining the principle on which they are based. This can best be achieved by means of a simple graphic representation.
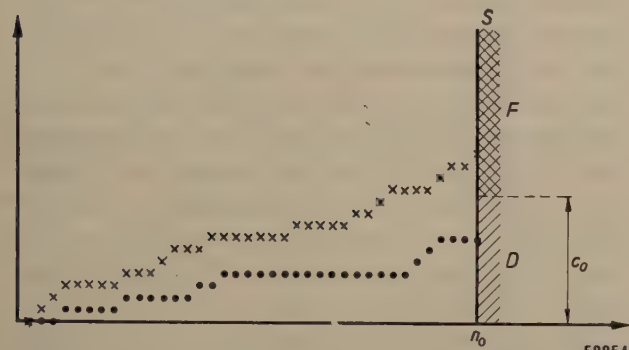
Fig. 2. The "random walk" diagram. The number of items inspected is plotted along the abscissa, the number of rejects observed among them along the ordinate. For a particular inspection lot the result of inspection is represented by the path indicated by the dots, for another lot by the crosses. The procedure of single sampling is illustrated by the screen S, erected in the point n_0. When the random walk ends in segment D the lot is accepted, when it ends in F the lot is rejected, c_0 being the acceptance number.

Along the abscissa of a set of rectangular axes we plot the total number of items inspected and along the ordinate the number of rejects found among them, see *fig. 2*. Starting from the origin we go in this diagram one step to the right for every good item inspected, and one step to the right and one step upward for every defective item. While inspection is going on we thus obtain step by step a detailed picture of the results obtained, as exemplified by the dots and crosses in fig. 2; such a path is called a "random walk", the entire graph being denoted as the "random walk diagram".

When several samples taken from the same lot are inspected the corresponding random walks will not, of course, be identical, but on the average they will run in the same direction. On the other hand, the greater the percentage defective of an inspection lot, the greater will be the average number of rejects in a sample drawn from it, and consequently the steeper the average slope of the corresponding random walks; thus the crosses in fig. 2 refer to a lot with a higher percentage defective than the dots.

This random walk diagram lends itself to a simple and easy illustration of the different sampling systems. Single sampling, for instance, is represented by erecting in the point n_0 on the abscissa a perpendicular divided into two segments D and F according to the acceptance number c_0. After inspecting a sample of size n_0 the random walk ends in some point on this perpendicular; if this point lies in D the lot is accepted, if in F it is rejected.

From single sampling we now pass on to "double sampling" by using two screens, S_1 and S_2, as in *fig. 3*, the first with an opening E in it. We now perform a sort of pre-selection. We start by taking a first sample of size n_1, which brings us to some point on S_1; if our random walk ends in D_1 the number of rejects is so low that the lot may at once be accepted, and if we end in F_1 the number of rejects is so high that the lot may be rejected. But if after the first sample we end somewhere in the opening E the case is considered as a doubtful one, so that it is desirable to collect some further data before deciding; we then take a second sample n_2 and base our final decision upon the total number of rejects in both samples together, according to the segments D_2 and F_2 of the second screen S_2.

Compared with single sampling, double sampling offers two advantages. Firstly many of us are inclined to give an inspection lot a second chance when the results of a first sample have not been entirely satisfactory, and this natural tendency is met by the double sampling principle. Secondly, decidedly good or decidedly bad lots are already brought to light by the first sample and this leads to a saving in the average number of observations. Against this we must set the drawback of a more
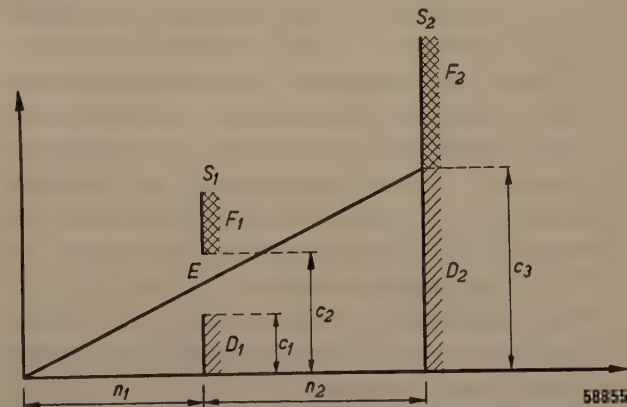

Fig. 3. The principle of double sampling depicted by two screens S_1 and S_2. S_1, corresponding to a sample size n_1, affords a pre-selection. Only those inspection lots for which the random walk after the first sample ends in the opening E, fixed by two criteria c_1 and c_3, are subjected to a further scrutiny, the result being considered as doubtful. This is done by taking a second sample of size n_2, and judging according to the screen S_2 and the criterion c_3.

complicated specification manifest from the larger number of data required: two sample sizes, n_1 and n_2, and three criteria c_1, c_2 and c_3 (see fig. 3).

Obviously by the same principles we might go farther still and sucessively develop three-fold, four-fold and in general m-fold sampling systems. Ultimately this leads automatically to the s e q u e n-t i a l   s a m p l i n g   s y s t e m which is represented by two parallel boundary lines dividing the random walk diagram into three regions D, E and F (*fig. 4*).
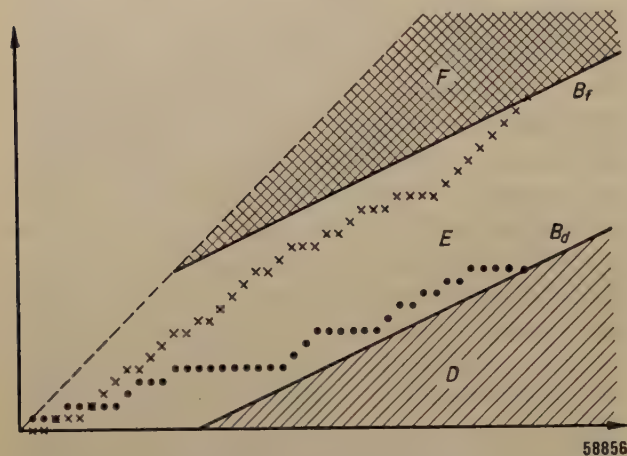
Fig. 4. The method of sequential sampling. When the random walk crosses the line B_d the lot is accepted, but when it crosses B_f the lot is rejected. (Since the number of rejects observed can never be greater than the number of units inspected, the random walk can never cross the line passing through the origin at an angle of 45°. This applies equally to figs 2 and 3.)

The regions D and F can be conceived as being the sum of an infinite number of screens such as S_1 in fig. 3. Inspection is continued as long as the random walks lie completely inside region E, but is stopped as soon as one of the two boundary lines B_d or B_f is crossed. If it is B_d the lot is accepted, if B_f it is rejected. When using the sequential method inspection is stopped as soon as sufficient data have been collected to reach a decision. This is the fundamental view from which sequential sampling has been developed, and from our above arguments it follows logically that by sequential sampling a still greater saving in the average number of observations can be obtained than with double or multiple sampling.

The denomination "s e q u e n t i a l" expresses that inspection is carried out sequentially, that is step by step; after each step it is decided whether the lot can be accepted or rejected, or whether we have to go on with inspecting another item. Another characteristic of this method is that the number of items to be examined is undetermined at the outset and depends on the results obtained while inspection is proceeding [2]).

In the foregoing the operating characteristic has been explained with reference to the single sampling system, but these characteristics can just as well be constructed for double and sequential systems, and this always produces curves of the same shape as those of fig. 1. It is possible to devise single, double and sequential sampling plans the characteristics of which are practically coincident. Such systems will on the average yield the same results as regards acceptance or rejection; they may for that reason be said to be "e q u i v a l e n t" in practical perform-ance, but this equivalency does not hold for the average size of the sample. On the contrary our above arguments lead us to expect that a double plan will require a smaller number of observations than its equivalent single sampling plan, this number being still lower in the case of sequential sampling. Thus a systematic comparison of equiv-alent sampling plans will enable us to decide to what extent the double or sequential system is of advantage; this principle will be worked out in detail in a subsequent article.

One might perhaps be inclined to conclude that sequential sampling, requiring the smallest number of observations, is always to be preferred, but this is by no means the case. Sequential sampling is less simple from an administrative point of view, and when the observations are relatively cheap this is a drawback that may well outweigh the advantage of a smaller number of data required. Moreover, in choosing a sampling plan most suitable for a practical problem many other factors play a part, as will now be discussed.

The practical choice of a sampling plan

In situations where lot-by-lot inspection may be practicable we can usually distinguish a "pro-ducer" and "consumer"; the consumer may be another factory, another department in the same works using the products delivered as components of a more complicated assembly, or a shopkeeper retailing to the public.

Both producer and consumer will lay down their requirements: the former demands that not too many "good" lots shall be rejected by the sampling inspection, while the latter demands

---

[2]) Double sampling was systematically applied for the first time by H.F. Dodge and H. G. Romig, whilst sequen-tial sampling was developed during the last war by A. Wald in the U.S.A. and by G. A. Barnard in England. See H. F. Dodge and H. G. Romig, Single and double sampling inspection tables, Bell system techn. J. **20**, 1-61, 1941; A. Wald, Sequential analysis, Wiley and Son, New York 1947; Ann. Math. Stat. **16**, 117-186, 1945. G. A. Barnard, J. Roy. Stat. Soc. Suppl. **8**, 1-27, 1946.

that not too many "bad" lots shall be accepted. In choosing a sampling plan attempts will be made to meet these somewhat opposing requirements.

Under favourable circumstances it may happen that all the lots produced are acceptable to the consumer, none being "bad", whereas under unfavourable conditions they may all happen to be "bad". In such cases lot-by-lot inspection is of no avail; the lots can either be delivered without any inspection at all, or they must all be examined for the full 100%. Consequently lot-by-lot inspection may be particularly useful where among a large number of good inspection lots we have reason to expect the occasional occurrence of a bad one. We are then in a position, by means of a sample, to discriminate good from bad and to pick out and correct the majority of the bad inspection lots in time.

Looked at in this way the results that can be achieved evidently depend on the frequency with which "bad" inspection lots occur; if no more than 1 in a 100 of the lots is bad the usefulness of lot-by-lot inspection will probably be less than when this ratio is 1 in 10. Another influencing factor is the difference between the "good" and the "bad" inspection lots. If this difference is relatively large it is comparatively easy to distinguish good from bad, but if the difference is only slight a satisfactory discrimination may require a sampling plan with a very steep operating characteristic needing very large samples. These are factors largely beyond our control, subject to changes owing to the continuous introduction of improvements in the production process, and which cannot easily be determined in a concrete way, but which are nevertheless of decisive importance in determining the proper choice of a sampling inspection procedure. As a rule we have to rely upon our "experience".

The cost of inspection is another point to be considered. If we are checking a diameter by means of a gauge or visually inspecting whether a product is burred or not, the cost of an observation is comparatively low and the size of a sample is of small concern. If, on the other hand, inspection is "destructive", as in testing a breakdown voltage or mechanical strength, every unit inspected is a loss and the number of items inspected must of necessity be limited; in such cases the use of double or sequential sampling may be advisable. Yet another situation is encountered in life tests where the use of a small sample is obviously desirable, but where sequential sampling must be ruled out since the inspection of one unit after another would take far too much time.

Finally it must be borne in mind that the object of detecting and correcting "bad" inspection lots is to avoid losses or damage that might result if they were passed on undetected. In this respect, too, practical circumstances may greatly diverge. Resistors or capacitors of inferior electrical quality lead to the production of badly functioning radio sets, and the tracing and correcting of the faulty elements involves considerable expense; only a very small percentage of such defects can be tolerated. But if the same components should lack, for example, one of the two terminal leads, this would be immediately noticed in the assembly shop and little harm is done. Similar considerations apply to products delivered to retailers. If among a consignment of incandescent lamps sold to the public the bulbs or a certain percentage of them are not in proper alignment with the base this will hardly be noticed, but if a high percentage should have a sub-normal life this will sooner or later be noticed and the customers will go elsewhere for their future supplies.

All in all we see that a number of economic factors play their part in the choice of a suitable sampling plan, factors that may differ widely from case to case, and the importance of which cannot be expressed in exact numerical values and which we must consequently evaluate on the basis of our practical experience.

From these considerations some important conclusions may be drawn as to the choice of a sampling plan. The introduction of a lot-by-lot inspection procedure can be regarded as a kind of insurance. The cost of the regularly recurring inspection constitutes the premium paid to cover ourselves against the losses resulting from failure to detect inferior products in time. But, as pointed out, we usually lack the precise data needed for an accurate evaluation of the premium to be paid, that is of the size of the sample. This at once implies that the finding of a suitable sampling plan is largely a question of experimental investigation. At the outset we have to go by the complaints concerning the quality of the goods supplied. Taking into account the consumer's requirements and the quality the producer claims as reasonably producable, we then make a preliminary choice in which we may be guided by our knowledge of the operating characteristics corresponding to various sampling plans. Practice will then have to show whether a wise choice has been made, this being judged by the degree in which complaints diminish or cease altogether. If such is not the case the sampling plan must be modified. The experience gained in

the foregoing experimental stage usually indicates the direction of the change required.

From the impossibility of determining the various economic factors with accuracy it also follows that the conditions to be satisfied in practice are not very stringent. If, for instance, a sampling plan possessing one of the operating characteristics of fig. 1 proves to be satisfactory, another plan with a characteristic deviating somewhat to the right or to the left, or being somewhat steeper or flatter, will be equally convenient. Theoretically there would, of course, be certain differences, but these will only become manifest from carefully kept records regarding the frequency and the nature of the complaints that still occur from time to time.

Taking a broad view of our problem we can now distinguish two different problems. First we have to specify the operating characteristic that is desired, then we have to construct a sampling plan satisfying this demand. These two problems cannot be completely separated one from the other, because in selecting a characteristic we have to take into account the number of observations required to obtain the curve selected. Nevertheless it is convenient to proceed in two stages as indicated.

Since no great precision is needed it has also been found sufficient to specify an operating characteristic by two numerical data (parameters), which simplifies our two problems to:

a) choosing the value of these parameters, and

b) finding a sampling plan corresponding to the values chosen.

In following this procedure it will be of great importance to adopt a set of parameters meeting the practical needs of the factory as far as possible. In this respect uniformity has not been reached, different sets of parameters having been introduced in the various sampling inspection tables so far developed. These sets of parameters and methods of constructing sampling plans when their values have been prescribed will form the subject of a subsequent article.

————

Summary. The quality of an inspection lot of mass-produced articles may often effectively be judged by means of a sample, the simplest method being that of single sampling characterized by a sample size n_0 and an acceptance number c_0 (number of rejects allowed in the sample). By plotting the probability of acceptance as a function of the percentage defective in the inspection lot the "operating characteristic" is obtained which is shown to represent the practical performance of a sampling method in a satisfactory manner. The principles of double and sequential sampling are next discussed, procedures which, though more complicated, possess the advantage of requiring a smaller number of observations than single sampling. It is indicated how the relative merits of these methods can be evaluated. Finally the factors, mainly economic, determining our choice of a sampling plan in particular cases are considered. For example the harm caused by "bad" lots not detected in time plays an important part. The choosing of a sampling plan is divided into two distinct stages; (a) the choice of an operating characteristic, and (b) the construction of a sampling plan possessing this characteristic. The latter problem, which is of a more purely mathematical nature, will form the main subject of a second paper.

THE FERRO-ELECTRICITY OF TITANATES

by G. H. JONKER and J. H. VAN SANTEN.

621.319.412.4

Some chemical compounds, among which Rochelle salt (potassium sodium tartrate) has been known the longest, when used as a dielectric between the plates of a capacitor show a remarkable behaviour similar to that of ferromagnetic materials. It has recently been found that certain titanates, of which barium titanate is an example, show this "ferro-electric" property. The behaviour of these latter substances can be explained to a certain extent in theory. Owing to the high values of their permittivities these titanates are employed in the construction of small capacitors of relatively large capacitance.

Introduction

Ferromagnetic materials such as iron, iron alloys and nickel, etc. are known to behave in the following way. At temperatures below a certain critical temperature T_c (the Curie temperature) these substances show a spontaneous magnetization. As the temperature rises this magnetization is reduced, first slowly and then gradually quicker until it becomes zero when $T = T_c$.

A given piece of ferromagnetic material usually contains a large number of small domains (Weiss domains) each with its own spontaneous magnetization. Since this magnetization may be directed in all sorts of ways, in the absence of an external magnetic field the total magnetization is as a rule zero or at least much smaller than what corresponds to the spontaneous magnetization of a single domain.

When an external magnetic field H is applied a reorientation of the spontaneous magnetization takes place in the various domains, as a consequence of which, even in a relatively weak field, the material as a whole suddenly shows a rather strong magnetization and thus assumes a considerable magnetic moment J per unit of volume. The relation between this magnetic moment J and the external field H is generally non-linear and not well defined. Starting from an originally non-magnetic sample of ferromagnetic material, as H increases from zero to high values there is usually at first a sharp increase and later on a less sharp increase of J until finally J reaches its ultimate saturation value J_s. As the value of H is reduced to zero again, J is also reduced, but the values of J with decreasing H are generally higher than those found with increasing value of H, so that when H has dropped to zero J shows a certain residual value (remanence) J_r *(fig. 1a)*.

If H is made to oscillate periodically between a high positive and a high negative value then the point representing the state of the material in an H-J diagram describes a so-called hysteresis loop (fig. 1b) [1]. On traversing the hysteresis loop once a conversion of energy into heat takes place in the material equal to $\oint H \mathrm{d}J$ per unit of volume (hysteresis losses).
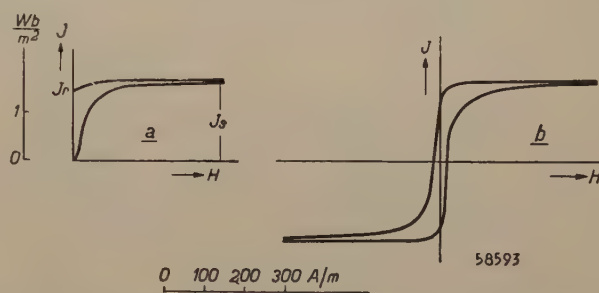

Fig. 1. Magnetic moment J per unit of volume as a function of the external field H for a ferromagnetic material (iron), a) static, b) in an alternating field (hysteresis loop).

Above the critical temperature T_c, which for pure iron for instance lies at 760 °C (1033 °K), the said materials are paramagnetic, that is to say, when placed in an external magnetic field H they assume a magnetic moment J per unit of volume proportional to H:

$$J = \varkappa H, \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

in which [2]

$$\varkappa = (\mu_r - 1)\, \mu_0 . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

The factor μ_r depends upon the temperature T

[1] See, for instance, J. J. Went, Philips Techn. Rev. **10**, 246-254, 1948, No. 8.

[2] For the sake of simplicity the material is assumed to be in the shape of a ring wound with wire (toroid). In such a form no complications arise owing to "demagnetizing forces". μ_r is the relative permeability with respect to vacuum; $\mu_0 = 4\pi/10^7$ henry/m. In this article the formulae are written according to the rationalized Giorgi system of electrical units with absolute volt and ampere. For further particulars about this system see P. Cornelius, Philips Techn. Rev. **10**, 79-86, 1948, No. 3 and Philips Res. Rep. **4**, 232-237, 1949, No. 3.

according to the equation:

$$\mu_r - 1 = \frac{A}{T - T_c}, (T > T_c) \quad \ldots \quad (3)$$

in which A represents a constant. It can be seen that μ_r increases sharply as T approaches the temperature T_c, so that T_c may indeed be regarded as a critical temperature.

be "para-electric". As a rule χ will be a function of the temperature. For most materials χ changes but little with T, but some are known where χ depends very strongly upon the temperature and shows at a certain temperature ϑ a high peak value. Some of these substances show at temperatures lower than ϑ a behaviour similar to that of ferromagnetic materials, in that P and E bear no longer
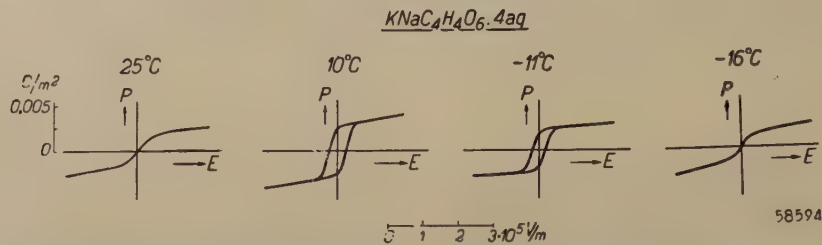
Fig. 2. The relation between P and E for Rochelle salt at different temperatures (taken from J. Hablützel, Helv. Phys. Acta 12, 489, 1939).
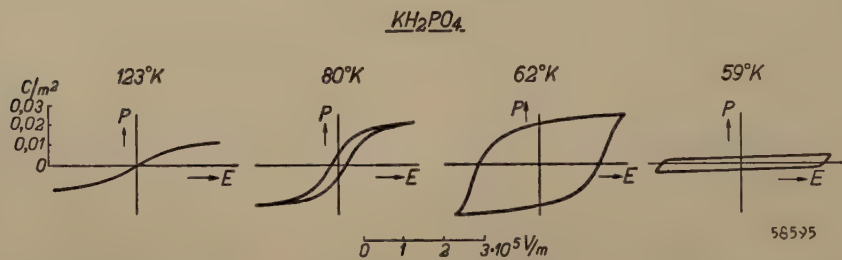
Fig. 3. The relation between P and E for KH_2PO_4 at different temperatures (taken from G. Busch and E. Ganz, Helv. Phys. Acta 15, 501, 1942).

Now a certain analogy exists between a magnetic material placed in an external magnetic field and a dielectric material placed in an external electric field.

In a flat capacitor, where the space between the plates is filled with a dielectric, there is usually a proportionality between the electric moment P per unit of volume and the field strength E, viz.
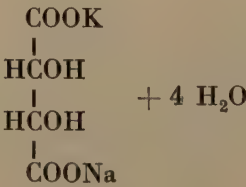
$$P = \chi E, \quad \ldots \ldots \ldots \quad (4)$$

in which [3])

$$\chi = (\varepsilon_r - 1) \varepsilon_0. \quad \ldots \ldots \ldots \quad (5)$$

Similar to the magnetic case, a medium which behaves according to (4), where χ denotes a constant independent of the field strength, could be said to

a linear relation to each other, whilst at the same time hysteresis occurs. Such materials are said to have an "electric Curie point" and are generally called ferro-electric. The material first (in 1918) found to possess this property is potassium sodium tartrate or Rochelle salt:

$$
\begin{array}{c}
\text{COOK} \\
| \\
\text{HCOH} \\
| \qquad + 4\ H_2O \\
\text{HCOH} \\
| \\
\text{COONa}
\end{array}
$$

This substance has a Curie point at $T = 300\ °K$ (27 °C), below which hysteresis occurs. Other materials showing a similar abnormal behaviour are KH_2PO_4 and allied compounds such as KH_2AsO_4, RbH_2PO_4, etc. In the case of KH_2PO_4 the Curie point lies at 124 °K. The hysteresis (P-E) diagrams for Rochelle salt and for KH_2PO_4 are given in figs 2 and 3 for a number of different temperatures.

---

[3]) ε_r is the so-called relative permittivity of the medium. For a vacuum the value of this constant is 1 and for other media mostly >1. The value of ε_0 is $10^7/4\pi c^2 = 8.85 \cdot 10^{-12}$ farad/m. The relation between P, E and the charge per unit of surface (electrical flux density) D is given by $D = \varepsilon_0 E + P$.

Fig. 2 shows that in the case of Rochelle salt at lower temperatures the non-linear relation between P and E and the hysteresis loop begin to disappear. It is said, therefore, that this material has not only the common or "upper" Curie point (27 °C) but also a "lower" Curie point (—19.5 °C). Something similar occurs with KH_2PO_4, see fig. 3, but here the case is somewhat different. At 58 °K the non-linear behaviour disappears but in this case the straight line indicating the relation between P and E has to be regarded as a hysteresis loop flattened in the vertical sense. We shall not, however, enter into these details, but will confine ourselves exclusively to the common or upper Curie point.

The causes of ferromagnetism and of ferro-electricity

The theory of magnetism is based on the conception that the electrons, owing to their describing orbits in the atoms and also rotating around their axis ("spinning"), are elementary magnetic dipoles. In the case of paramagnetism these elementary magnets (for the sake of simplicity denoted as spin vectors) are orientated by an external field, whilst the thermal movement tends to disturb this order. The spontaneous magnetization in the case of ferromagnetism is brought about by a gain in energy obtained when all the spin vectors are directed parallel to each other. This is not due to the magnetic interaction between the elementary magnets (it is too small for that) but arises from an additional gain in energy which can only be accounted for by quantum mechanics. According to this theory, which we cannot go into here, it can be understood more or less why it is that this additional interaction is sufficient to lead to ferromagnetism in the case of some materials and not so with others.

The behaviour of Rochelle salt must likewise be ascribed to dipoles present in the material, though in this case these are of an electrical nature. These electrical dipoles arise not from the properties of the electrons but from the arrangement of the ions in the crystal. The crystal lattice of Rochelle salt contains H^+-ions placed asymmetrically between negative oxygen ions. This results in the presence of electric dipoles. Now these dipoles have the tendency to orientate each other and in this case the energy of the dipole forces is indeed sufficient to account for this orientating effect. Thus it is that the spontaneous polarization in this case is of a nature different from the magnetization in the magnetic case.

Barium titanate and allied compounds

We shall now discuss a group of compounds whose behaviour corresponds in many respects to that of Rochelle salt but is easier to discuss because these compounds have a very simple (cubic) crystal structure. In the years 1940-1945 it was found [4] that compounds of TiO_2 with oxides of bivalent metals, for instance $BaTiO_3$ and $PbTiO_3$ and mixed crystals of these, show as a function of the temperature a high peak in the value of ε_r, whilst other similar compounds such as $SrTiO_3$ and $CaTiO_3$, as well as TiO_2, show a sharp increase of ε_r with decreasing temperature as far as the absolute zero point [5].

The titanates referred to above are prepared by ceramic methods, mixtures of TiO_2 and, for instance, $BaCO_3$ being ground together, moulded to the desired shape and fired at 1300-1400 °C. X-ray crystallographic examination shows that a complete reaction takes place. The preparation of $PbTiO_3$ is somewhat more difficult because in the process of sintering some of the PbO evaporates. It is due to the great stability of these ceramic materials that use can be made of their very high permittivities for practical purposes, whereas this is not easily possible with Rochelle salt and allied materials. Something more will be said about this at the end of this paper.

The permittivity of titanates as a function of temperature

The behaviour of the permittivity of $BaTiO_3$ as a function of T is rather complex. As is to be seen from fig. 4, ε_r for $BaTiO_3$ has a very high maximum ($\varepsilon_r \approx 10,000$) at $T = 396$ °K (123 °C) and two smaller maxima in the neighbourhood of
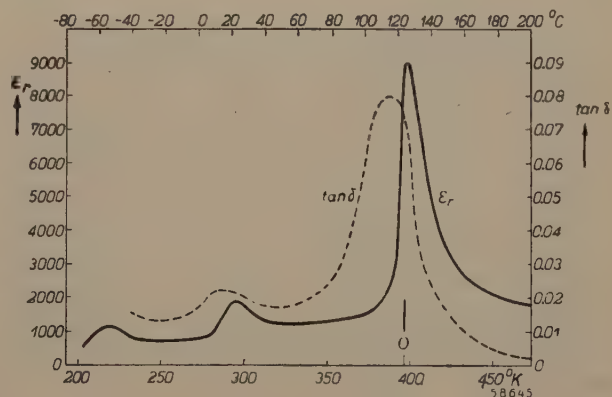

Fig. 4. The permittivity ε_r and the power factor $\tan \delta$ of $BaTiO_3$ as functions of the temperature, measured at a low field strength (1000 V/m).

[4] B. Wul, Dielectric constant of some titanates, Nature 156, 480, 1945; Willis Jackson and W. Reddish, High permittivity crystalline aggregates, Nature 156, 717, 1945; E. Wainer, High titania dielectrics, Trans. Electrochem. Soc. 89, 331-356, 1946; A. von Hippel. R. G. Breckenridge, F. G. Chesley and L. Tisza, High dielectric constant ceramics, Indust. Eng. Chem. 38, 1097-1109, 1946.
[5] E. J. W. Verwey and R. D. Bügel, Philips Techn. Rev. 10, 231-238, 1948, No. 8.

$T = 10\ °C$ and $-70\ °C$. The main maximum is greatly affected, as far as its height and width are concerned, by the degree of sintering and the purity of the material. The temperature ϑ of the main maximum forms the boundary between two temperature regions in which the electrical and other properties are very different.

In the range $T > \vartheta$ the temperature coefficient of the permittivity is negative. In this range ε_r can be represented as a function of T by the simple formula

$$\varepsilon_r = \frac{A'}{T - \vartheta}, \quad \ldots \ldots \ldots \quad (6)$$

in which A' is a constant. This formula (6) is entirely analogous to formula (3) which applies for the magnetic case. (The fact that in formula (6) we have ε_r instead of $\varepsilon_r -1$ is of no importance because we are here considering dielectrics of which $\varepsilon_r \gg 1$.)

The other titanates and TiO_2 show the same behaviour in a certain range of temperature; the value of the constant A' is always about 1.0×10^5, but the value of ϑ differs from case to case, being 396 °K for $BaTiO_3$, 800 °K for $PbTiO_3$ and about 0 °K for $SrTiO_3$. A similar formula applies for $CaTiO_3$ and TiO_2 but with a negative value of ϑ. Since T is always positive, throughout the whole temperature range $(T > 0\ °K)$ $SrTiO_3$, $CaTiO_3$ and TiO_2 therefore behave in the same way as $BaTiO_3$ above 396 °K.

All the above titanates have the same crystal structure. They can be used for producing mixed crystals in all proportions by sintering suitable mixtures together. An examination for instance of mixed crystals of $BaTiO_3$ and $SrTiO_3$ shows that a practically linear relation exists between the mixing proportions and the peak temperature ϑ (fig. 5).

Thus by starting with mixtures of $BaTiO_3$ and $SrTiO_3$ the peak temperature can be made to lie anywhere between 396 ° K and 0 °K, so that it is possible, for instance, to study and apply the properties in the neighbourhood of $T = \vartheta$ at room temperature.

Formula (6) holds equally well for the mixed crystals. The constant A' has practically the same value also in this case.

Obviously, near the maximum of ε_r $(T = \vartheta)$ formula (6) no longer holds because it would lead to the value $\varepsilon_r = \infty$. But within the range from a few degrees above $T = \vartheta$ up to the highest temperatures to be considered formula (6) does indeed hold good. Provided that T is chosen sufficiently high above ϑ then in the range of $T > \vartheta$ the

dielectric losses are fairly small. The temperature coefficient of the dielectric constant is more strongly negative as T approaches ϑ, and there the losses are greatly increased.
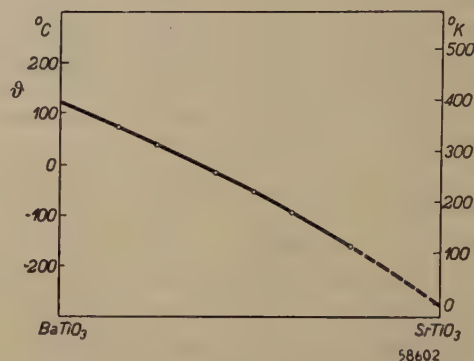

Fig. 5. Peak temperature ϑ as function of the proportions of $BaTiO_3$ and $SrTiO_3$ in mixed crystals.

Electrical behaviour for $T < \vartheta$.

For the titanates having a positive value of ϑ and thus a permittivity showing a high maximum (see fig. 4) the dielectric properties in the range $T < \vartheta$ are entirely different from those in the range $T > \vartheta$. The temperature coefficient in the range $T < \vartheta$ may be positive or negative, and, taken absolutely, generally smaller than in the range $T > \vartheta$. In the range $T < \vartheta$ the losses are considerable. This is connected with the fact that there appears to be a non-linear relation between P and E and hysteresis is found to occur. Because of these facts, in this temperature range these substances strongly resemble ferromagnetic materials and Rochelle salt.

When the material is used as a dielectric between the plates of a capacitor under the conditions mentioned above the charge Q of the capacitor is no longer proportional to the potential difference V between the electrodes. Instead of a simple
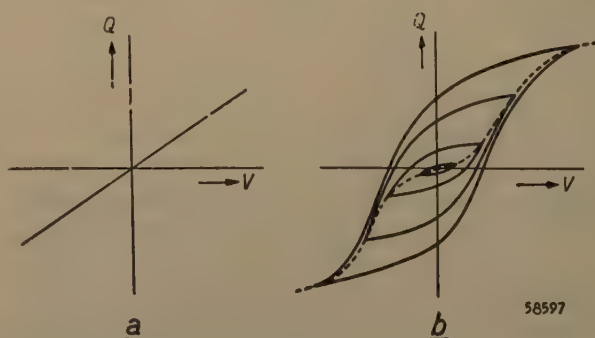

Fig. 6. The charge Q of a capacitor with $BaTiO_3$ as dielectric, as a function of the potential difference V, a) in the range $T > \vartheta$, b) in the range $T < \vartheta$. The dotted line represents (schematically) the initial curve.

Q-V figure [6]) as represented in *fig. 6a* we then get a hysteresis loop (fig. 6b).

It is not easy to determine the hysteresis figure statically because the charges arising, in so far as they are present at the crystal surface, mostly leak away quickly. Measurements are therefore usually taken with the aid of alternating voltages of different amplitudes. The initial curve can be determined by assuming that the ends of all hysteresis loops lie on this curve (see the dotted line in fig. 6b). It is most easily determined by the method to be mentioned farther on, where use is made of a cathode-ray oscillograph.

The initial curve of a titanate is drawn separately in *fig. 7*. The initial slope is of the same order as the slope found with the highest attainable voltages. The charge and thus the polarization of the dielectric can be regarded as consisting of a part linearly dependent upon the field strength and another part which becomes saturated at a certain field strength.
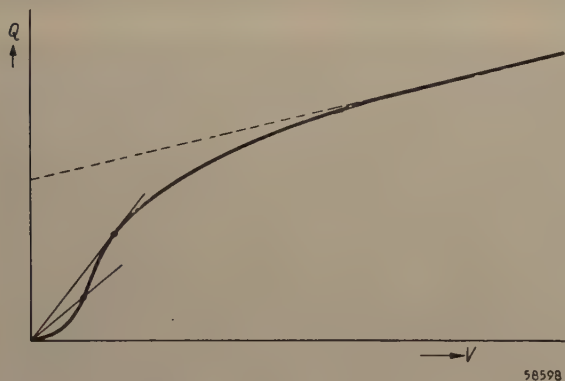

Fig. 7. The initial curve of a titanate (schematically).

As a result of the contribution yielded by the part that becomes saturated, the effective permittivity, determined by the ratio of the amplitudes of the charge and voltage, first rises sharply with increasing amplitude of the alternating voltage and then gradually drops until the ultimate value is of the same order as the initial value (*fig. 8*).

In contrast to the effective permittivity the reversible permittivity (slope of a small hysteresis loop starting from an arbitrary working point) is relatively independent upon the circumstances. It is approximately equal to the initial value of the permittivity (slope of a small hysteresis loop described around the origin, $Q = 0$, $V = 0$).

---
[6]) The potential difference V is proportional to E and the charge Q is proportional to D, thus to ε_r. P is proportional to $\varepsilon_r - 1$ (see footnote [3])) but in our case invariably $\varepsilon_r \gg 1$. Therefore the Q-V diagram is practically equal to the P-E diagram.

As is already to be deduced from fig. 6b, in the range $T < \vartheta$ the ε_r-T curve of fig. 4 is dependent upon the alternating voltage used. This curve represents the initial value and applies for a small field strength (< 1000 V/m). When measurements are carried out at higher voltages we get the effective permittivity and thus the part of the curve in the range $T < \vartheta$ changes considerably (*fig. 9*), whilst also the losses are greatly increased.
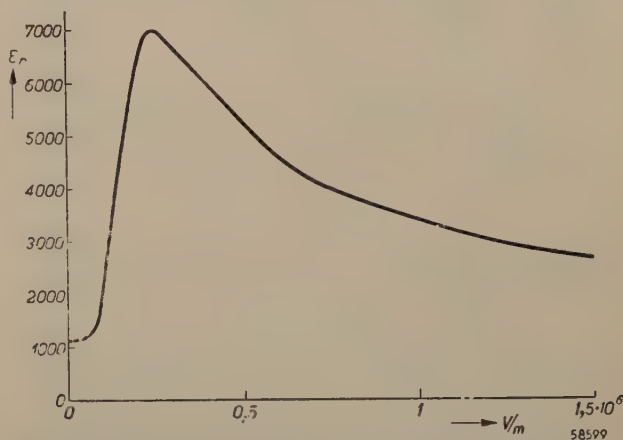

Fig. 8. The effective permittivity of a titanate mixed crystal ((87.5 % Ba, 12.5 % Sr) TiO$_3$, $T = 20$ °C) as a function of the amplitude of the alternating voltage.

From this it follows that at lower temperatures a higher field strength is needed to reach the maximum ε_r. This is to be seen also from *fig. 10*, where the initial curve has been drawn for two temperatures.
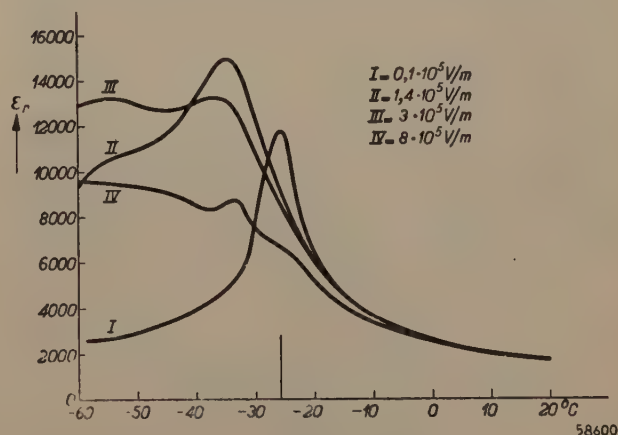

Fig. 9. The initial value (curve *I*) and the effective value ε_{eff} of the permittivity for different amplitudes of the alternating voltage as functions of T (for the mixed crystal (55 % Ba, 45 % Sr) TiO$_3$).

Method of investigating the dielectric behaviour

This investigation is made largely with the aid of a cathode-ray oscillograph, by means of which pictures can be obtained of the hysteresis loop. As regards the circuit reference is made to *fig. 11* and the explanation given in the caption.

The capacitor C_1 (fig. 11) containing the material to be investigated is placed in a bath of paraffin that can be heated to different temperatures. In this way it can be clearly seen how the hysteresis
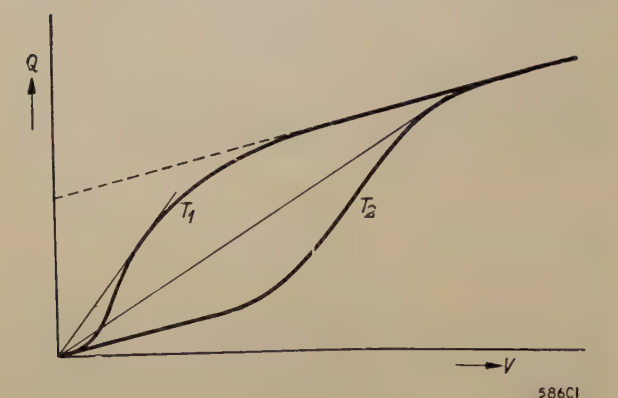

Fig. 10. The initial curve of a titanate (schematically) at two temperatures $(T_1 > T_2)$.

loop changes into a linear characteristic at $T = \vartheta$. Photographic recordings of this are given in *fig. 12*. In *fig. 13*, in addition to a voltage of 50 c/s also a small voltage with a frequency which is a multiple of 50 c/s (in the case in question 1750 c/s) is applied to the plates of the capacitor, making the reversible permittivity visible. At the point of intersection of the axes there is a loop with a very small amplitude, the slope of which corresponds to the initial value of the permittivity.

Fig. 13 clearly shows that there is little difference between the slope of the initial loop, that of the reversible loops and the final slope of the large hysteresis loop.

Explanation of the exceptional behaviour of titanates

What are the causes of the remarkable behaviour of the dielectric constant of titanates? Let us first
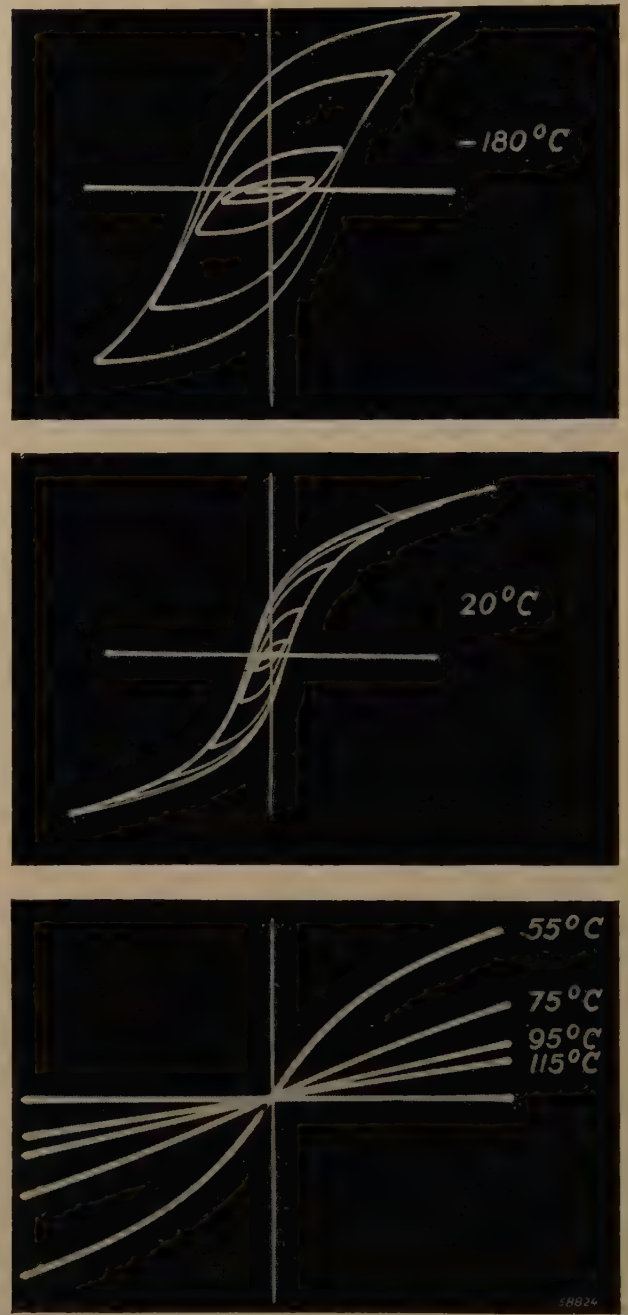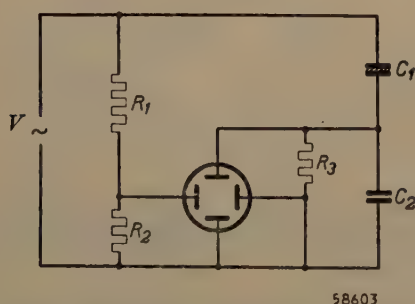

Fig. 11. Circuit diagram for investigating electric hysteresis (C. B. Sawyer and C. H. Tower, Phys. Rev. **35**, 269, 1939). The horizontal deflection voltage is directly proportional to the total voltage V and thus practically also proportional to the voltage at the capacitor C_1 with the substance under examination, since $C_2 \gg C_1$ (C_2 is an auxiliary capacitor with normal dielectric). The vertical deflection voltage V_2 is proportional to the charge of C_2 and thus also proportional to the charge of C_1, since $Q_2 = Q_1$.

consider the region $T > \vartheta$. According to Clausius-Mosotti and Lorentz:

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{1}{3\varepsilon_0} N\alpha \quad\quad . . . . . . (7)$$

where N represents the number of titanate "molecules" per unit of volume and α the sum of the

Fig. 12. Photographs of the Q-V diagram of a mixed crystal (79 % Ba, 21 % Sr) TiO_3 at different temperatures and with different amplitudes of voltage, recorded with the cathode-ray oscillograph. At -180 °C the hysteresis loop is very broad, at 20 °C narrower, whilst at the transitional point $\vartheta \approx 55$ °C the two branches practically coincide, a straight characteristic being obtained at a somewhat higher temperature.
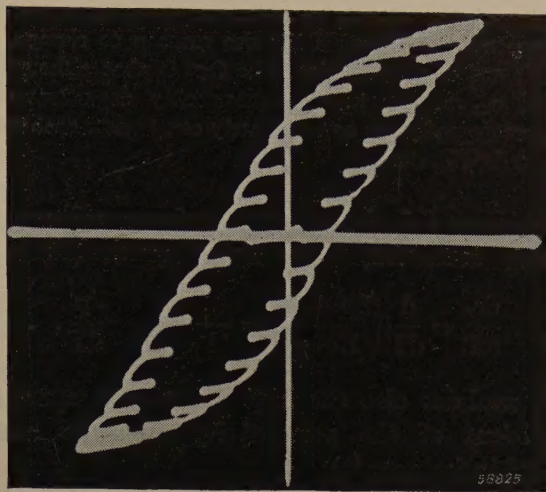
Fig. 13. Photograph of a hysteresis loop of $BaTiO_3$ with parasitic loops the slope of which denotes the reversible permittivity; furthermore, at the point of intersection of the axes a small loop is to be seen showing the initial value of the permittivity.

polarizabilities of all the ions in the molecule [7]. Substituting p for the term $(1/3 \, \varepsilon_0)Na$ we get

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = p, \quad \text{or} \quad \varepsilon_r = \frac{2p + 1}{1 - p}. \quad \ldots \ (8)$$

For most solids the values of p are between 0.3 and 0.9. If $p = 1$ the ε_r would be infinite. Now the abnormal behaviour of titanates is to be accounted for by the fact that for these materials the value of p happens to differ but little from unity.

In the article referred to in footnote [5] it has already been pointed out that the quantity p can be considered as being built up from various contributions according to the formulae

$$a = a_e + a_i + \ldots, \quad \ldots \ldots \ (9a)$$

$$p = p_e + p_i + \ldots, \quad \ldots \ldots \ (9b)$$

in which $p_e = (\frac{1}{3} \, \varepsilon_0)Na_e$ is the contribution of the electrons and $p_i = (\frac{1}{3} \, \varepsilon_0)Na_i$ that of the ions.

In addition to the contribution of a_e towards a, resulting from the deformation of the atomic electron clouds under the influence of an electric field (electronic polarization), and the contribution a_i which follows from the relative displacement of the ions in a crystal under the influence of an external field, account has also to be taken of the possibility of a contribution towards the polarization due to the presence of fixed dipoles in the crystal which are orientated under the influence of the field. As will be explained farther on, in the range $T < \vartheta$ it is to be assumed that dipoles are present in $BaTiO_3$. To account for the behaviour of titanates in the range $T > \vartheta$, however, there is *no* need to assume the presence of permanent dipoles, contrary to the case with Rochelle

salt, where dipoles play a part both in the range of $T < \vartheta$ and in that of $T > \vartheta$.

Another important point is the following: in addition to the contributions towards a derived from the electron and from the ions, in the case of titanates, even disregarding the presence of permanent dipoles, there is an additional contribution due to the fact that although the structure of the lattice is cubic there are still some ions in the lattice (O^{2-} ions) which are not cubic-symmetrically surrounded by other ions. It can be proved [8]) that this results in an increase of a and thus leads to an additional increase of the value of p, so that one may write $a = a_e + a_i + a_{as}$, in which a_{as} represents the correction term for the asymmetrical environment, and thus also $p = p_e + p_i + p_{as}$.

The contribution p_e can be calculated separately, since according to Lorentz

$$p_e = \frac{n_0^2 - 1}{n_0^2 + 2}, \quad \ldots \ldots \ (10)$$

where n_0 represents the optical index of refraction extrapolated to infinitely low frequencies. For $BaTiO_3$, for instance, this contribution towards p has the value 0.63, so that a contribution of $p_i + p_{as}$ of 0.37 is required to give $p = 1$.

The high index of refraction is related to the high density (large number of easily polarizable O^{2-} ions in a small volume) and to the fact that in the case of $BaTiO_3$ and allied compounds the crystal absorption (responsible for refraction and dispersion) takes place at fairly low frequencies.

In order to understand the high value of p_i, the contribution due to the displacement of ions under the influence of an external electric field, it is necessary to investigate the crystal structure of the titanates more deeply.

Crystal structure of titanates

The compounds $CaTiO_3$, $SrTiO_3$, $BaTiO_3$, $PbTiO_3$ show the perovskite structure, so-called after the mineral perovskite $CaTiO_3$ (see *fig. 14*). The cubic
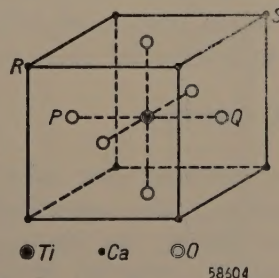

Fig. 14. Unit cell of perovskite ($CaTiO_3$).

[7]) The formula written in this way applies when Giorgi units are employed ($a \ldots$ ($A \cdot sec \cdot m^2 / V$)). In cgs units we get on the right-hand side $(4\pi/3) \, Na \, (a \ldots (cm^3))$.

[8]) G. H. Jonker and J. H. van Santen, Properties of barium titanate in connection with its crystal structure, Science **109**, 632-635, 1949. See also J. H. van Santen and W. Opechowski, Physica **14**, 545-552, 1948, No. 10.

unit cell has Ca-ions at the corners, O-ions at the face-centre and a Ti-ion at the cube centre.

Goldschmidt [9]) has already made an exhaustive study of this structure and found that compounds with the formula ABO_3 have to satisfy a certain condition to be able to crystallize in the perovskite lattice. This condition is that the ions must have suitable dimensions, which in geometrical terms means that the length of the unit cell calculated from the radii of the ions along the line PQ must equal the length calculated from the diagonal RS; in other words

$$r_A + r_O = (r_B + r_O)\sqrt{2},$$

in which O represents the oxygen ion, B the titanium ion and A the alkaline-earth ion.

Goldschmidt describes this condition with a parameter

$$t = \frac{r_A + r_O}{(r_B + r_O)\sqrt{2}},$$

which for the perovskite structure has to be approximately equal to 1. This is not a stringent requirement; among the known perovskites t varies between 0.80 and 1.14. If, however, t is less than 0.80 or greater than 1.14 the compounds ABO_3 crystallize according to a different type of crystal.

If t answers the requirement $0.8 < t < 1.14$ the structure may still differ from the purely cubic type of fig. 14 owing to a small deformation of the lattice (tetragonal and monoclinic perovskite). Nevertheless it appears that at least in the tetragonal case at sufficiently high temperatures the structure becomes purely cubic. This is for instance the case with $BaTiO_3$ for $T > 396\,°K$. Since $t > 1$, which means that $r_{Ba} + r_O > (r_{Ti} + r_O)\sqrt{2}$, the Ti-ion has at its disposal a large space in the crystal. It is therefore easily displaced and is capable of yielding a larger contribution towards p_i than is the case with other titanates. There are a number of mixed crystals of $BaTiO_3$ with other perovskites which have the same property.

The point of transition

Let us now consider what takes place at the point of transition, taking again $BaTiO_3$ at a temperature above $396\,°K$. Let, for instance, p equal 0.99 ($\varepsilon_r \approx 3000$, $T = 420\,°K$). When the temperature decreases the density of the crystal and thus N will increase, and with that also p.

Although theoretically possible, any change in a with the temperature, which would likewise cause p to change, can be left out of consideration. The fact is that although a_e, a_i and a_{as} each in itself is presumably dependent upon T, these dependencies approximately compensate each other, so that $da/dT \approx 0$.

Owing to this slight increase of p, ε_r will now be greatly increased, since according to (8)

$$\frac{d\varepsilon_r}{dT} = \frac{d}{dT}\left(\frac{2p+1}{1-p}\right) = \frac{3}{(1-p)^2} \cdot \frac{dp}{dT}. \quad . \ (11)$$

This explains also the high negative temperature coefficient of the permittivity (see the article quoted in footnote [5])).

With increasing ε_r a smaller external field is required to bring about a certain polarization of the crystal. Where $p = 1$ even an infinitesimally small external field is sufficient to cause a finite polarization. This becomes manifest in the fact that the state of the crystal is then unstable; in other words, when $p = 1$ the cubic arrangement from which we started is no longer the most stable state. Spontaneous polarization takes place, accompanied by the formation of dipoles (with simultaneous displacement of all Ti-ions) and a small deformation of the originally cubic unit cell.

Thus the temperature at which $p = 1$ represents a crystallographic point of transition. Below this temperature $BaTiO_3$ does indeed still retain its perovskite character but the unit cell is then no longer purely cubic, being slightly distorted tetragonally and assuming the shape of a four-sided prism. This is also apparent from fig. 15, where the axial ratio of the unit cell is represented as a function of T.
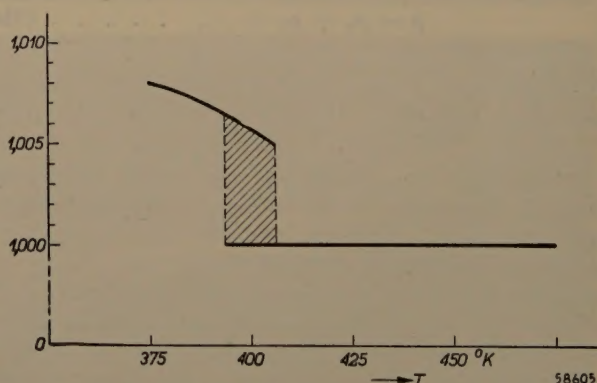

Fig. 15. Variation of the axial ratio of $BaTiO_3$ (according to measurements taken by M. G. Harwood, P. Popper and D. F. Rushman, Nature **160**, 58, 1947).

It appears that in this transition an originally cubic monocrystal mostly changes into an agglomerate of tetragonal domains. Each of these domains is spontaneously polarized.

[9]) See for instance, H. D. Megaw, Proc. Phys. Soc. **58**, 133-152, 1946.

The cause of hysteresis

In a crystal in which the spontaneously polarized domains described above occur there will be no outward evidence of an electric field, because according to the laws of probability the direction of the polarization will be distributed over the various possibilities. Upon an electric field being applied not only will the normal mechanism of polarization come into play but there will also be an expansion of some domains at the cost of others, thus giving rise to a volume polarization. This polarization continues to exist when the external field is removed, so that the phenomenon of remanence arises and in the case of alternating fields a hysteresis loop is described.

A crystal treated in this way may, therefore, show a macroscopic moment; in other words it may be an "electret". Actually anything of this nature will not be noticeable for long because charges will very soon flow along the surface and compensate the field of the internal charges. Inside the crystal, however, the state of polarization may indeed continue to exist. One of its results is the occurrence of a so-called induced piezoelectric effect [10]).

Since the area $\oint EdP$ of the hysteresis loop is proportional to the work performed by the electric forces there will obviously be great losses in the temperature range in which hysteresis occurs, particularly if the material is exposed to a strong alternating field.

Application in capacitors

In the application of $BaTiO_3$ and mixtures thereof with other titanates in capacitors, with a view to utilizing the high value of ε_r, the great dependence of the dielectric constant upon the temperature is found to be a drawback. In pure $BaTiO_3$ ε_r is only fairly constant between 30 °C and 80 °C, having a value of about 1200. What is desired, however, is to utilize the very much higher values reached round about the transition temperature; as stated above; by using mixed crystals of titanates this transition point can be made to lie at practically any temperature desired.

The fact still remains however that the very high values of ε_r desired occur only in a small temperature range and that within that range ε_r is greatly influenced by the temperature.

This can be remedied in the first place by admixing foreign materials to the titanate mixtures.

The width of the peak, which in pure titanates is in the order of 20-30 °C, can then be raised to 40-50 °C, though at a slight cost of the height of the maximum.

The range within which ε_r has a high value can be further extended by combining two suitable kinds of mixed crystals with different proportions of $SrTiO_3$ and $BaTiO_3$, each placed as a dielectric in a separate capacitor. The most economical way is to connect these capacitors in parallel. It depends entirely upon the requirements to be made of such a composite capacitor as to how far the peaks of the two mixed crystals are to be separated (*fig. 16*).
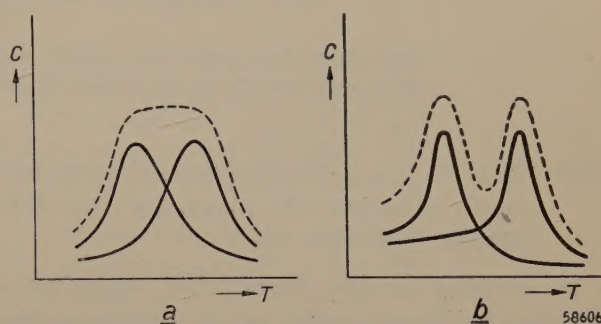

Fig. 16. Schematic representation of the result of capacitors with different peak temperatures being connected in parallel. *a*) Fairly wide peaks close together, *b*) narrower peaks farther apart.

In practice one may choose the tubular type commonly used for ceramic capacitors, where the tube is made partly of one mixture and partly of the other (see *fig. 17*). Of course one cannot then avoid having to use titanates in a temperature range in which the hysteresis phenomena and high
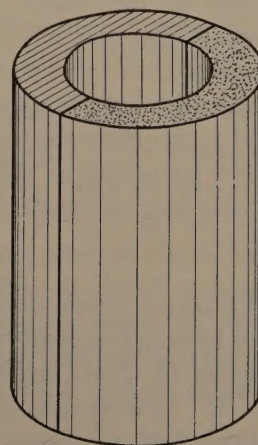
Fig. 17. Filling of a tubular capacitor in which two mixtures of titanates are placed parallel to each other.

---

[10]) See for instance J. C. B. M i s s e l, Philips Techn. Rev. 11, 145-150, 1949, No. 5.

losses occur. Consequently the possibilities of application of these capacitors are somewhat limited and they are not suitable for use in tuned circuits, though they can be used, for instance, as coupling capacitors.

———

Summary: The most important properties of ferromagnetic materials are briefly discussed in connection with the fact that in an electric field the behaviour of certain chemical compounds such as $KNaC_4O_6$ $4H_2O$. (Rochelle salt) and KH_2PO_4 very closely resembles the phenomenon of ferromagnetism. Particularly of importance is the high value of ε_r, similar to that of μ_r in the case of ferromagnetic materials. Further, in analogy with ferromagnetic materials, below a certain critical temperature ε_r becomes dependent upon the field strength and hysteresis occurs. Such is found to be the behaviour of barium titanate and allied compounds. In the latter case, owing to the comparatively simple crystal structure, these phenomena are easily explained theoretically. The theoretical explanation of the behaviour of titanates is given in broad lines, whilst at the same time it is shown how the very high value of the permittivity found with these materials can be utilized for the construction of capacitors required to possess a fairly constant and large capacitance within a certain temperature range, whilst being of the smallest possible dimensions.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF THE N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk can be obtained free of charge upon application from the address printed on the back cover.

1854: K. F. Niessen: Nodal planes in a perturbed cavity resonator, III (Appl. sci. Res. B 1, 284-298, 1949, No. 4).

In choosing in the unperturbed resonator a vibration having two nodal planes (one perpendicular and one parallel to the movable wall) the mathematics required is of quite another nature than in Part II. Again a combination of functions is chosen, the coefficients of which have to be determined. Now it appears that all these coefficients, except that belonging to the original function, must be of the order of δ, just as was found in Part I. There is only one possible solution and its nodal figure is permanent, being no function of time. Instead of the two intersecting nodal planes $y = a/2$ and $z = a/2$, existing for $\delta = 0$, we obtain for $\delta \geqslant 0$ two sheets, the intersection of which with a plane perpendicular to the x-axis resembles an orthogonal hyperbola and the difference between the new and the original nodal system is relatively large in the neighbourhood of the central axis of the resonator. The use of this effect for the construction of special resonators containing a "side room" and a bar is shown. (See also abstracts Nos 1815 and 1853.)

1855: Th. J. Weijers: Toelaatbare onderlinge storing van twee in frequentie gemoduleerde omroepzenders (T. Ned. Radiog. 14, 61-72, 1949, No. 3). (Permissible mutual interference of two F.M. broadcast transmitters; in Dutch.)

In order to obtain data about the selectivity of an f.m. broadcast receiver, the measurements have to be made according to the two-signal method. The results of these measurements, taken on three different receivers, are given and discussed. An interesting result is that, for an f.m. broadcast service in a given area, the required frequency band is reduced considerably if, by repeating the same frequency band at a sufficient distance, the central frequencies of the second set of transmitters are shifted over 50 kc/s or more.

1856: A. Bril: On the saturation of fluorescence with cathode-ray excitation (Physica 's Grav. 15, 361-379, 1949, No. 3/4).

Measurements of the light output as a function of the beam current under stationary cathode-ray excitation are performed for zinc silicate activated with various concentrations of Mn or Ti and for zinc sulphide activated with ZnCl and with AgCl. The intensity of fluorescence saturates at high current densities. Under continuous electron bombardment the saturation depends on the activator content, on the presence of quenchers and on the decay time of the fluorescence. The concentration of centers in ZnS-ZnCl is estimated. In the case of a spot of high current density scanning a definite area, the saturation does not depend on the decay time of fluorescence if the time in which the electron beam passes a certain point of this area is small with respect to the decay time.